

Lehrbuch der psychologischen Diagnostik

Lehrbuch der psychologischen Diagnostik

Mit Hinweisen zur Intervention

von

Hermann-Josef Fisseni

2., überarbeitete und erweiterte Auflage



Hogrefe · Verlag für Psychologie
Göttingen · Bern · Toronto · Seattle

Prof. Dr. Hermann-Josef Fisseni, geb. 1932. Von 1955 bis 1963 Studium der Philosophie, Theologie und Psychologie. 1969 Diplompsychologe; 1973 Promotion bei Prof. Dr. H. Thomae und Prof. Dr. Ursula Lehr, Bonn. 1979 Habilitation. Seit 1982 Professor für Psychologie in Bonn mit den Arbeitsschwerpunkten Psychodiagnostik und Persönlichkeitspsychologie.

Die Deutsche Bibliothek - CIP-Einheitsaufnahme

Fisseni, Hermann-Josef

Lehrbuch der psychologischen Diagnostik : mit Hinweisen zur Intervention/ von Hermann-Josef Fisseni. - 2., überarb. und erw. Aufl. - Göttingen ; Bern ; Toronto ; Seattle : Hogrefe, Verl. für Psychologie 1997
ISBN 3-8017-0982-5

© by Hogrefe-Verlag, Göttingen . Bern • Toronto • Seattle 1990 und 1997
Rohnsweg 25, D-37085 Göttingen



Das Werk einschließlich aller seiner Teile ist urheberrechtlich geschützt. Jede Verwertung außerhalb der engen Grenzen des Urheberrechtsgesetzes ist ohne Zustimmung des Verlages unzulässig und strafbar. Das gilt insbesondere für Vervielfältigungen, Übersetzungen, Mikroverfilmungen und die Einspeicherung und Verarbeitung in elektronischen Systemen.

Satz: Druckvorlagen Bernert, Göttingen
Druck: Dieterichsche Universitätsbuchdruckerei
W. Fr. Kaestner GmbH & Co. KG, D-37124 Rosdorf
Printed in Germany
Auf säurefreiem Papier gedruckt

ISBN 3-8017-0982-5

Vorwort

Dieses Lehrbuch wurde für „Studierende“ geschrieben, welche das Fach „Psychologische Diagnostik und Intervention“ kennenlernen wollen.

Darum vermittelt das Buch Kenntnisse und Techniken, mit denen jemand vertraut sein sollte, wenn er in der Praxis „Psychologie anwenden“ will. Das Buch kann nur ein Fundament legen - der Studierende muß diese Grundlage erweitern; das Nachwort bietet dafür Ratschläge an.

Die Gliederung des Buches wurde konzipiert von der „diagnostischen Situation“, also von der Anwendung her. In Diagnostik wie in Intervention muß der „Anwender“ Kenntnisse abrufen, die sich beziehen auf Regeln und Gesetze der Datensammlung und Datenintegration, auf Kenntnisse von Einzelverfahren, die er nur handhaben kann, wenn er die Konzepte versteht, nach denen sie entworfen wurden.

Die Kenntnisse von Diagnostik und Intervention stellen wir uns als gestaffelt vor:

- Darum umreißt das Buch *zuerst* Herkunft, Eigenart und Aufgabenfelder der Anwendungsmethodologie Diagnostik und Intervention, charakterisiert ihren sozialen, finalen und ethischen Kontext (Teil I).
- *Sodann* behandelt es in drei Durchgängen größere Lehreinheiten: Grundkenntnisse, spezielle Einzelverfahren und spezielle Einzelfragen (Teil II bis IV).
- *Schließlich* stellen wir Beispiele integrativer multimodaler Diagnostik und Intervention vor (Teil V).

Die gesamte Darstellung ist stärker an Individualdiagnostik als an Reihenuntersuchungen oder Forschungsaufgaben orientiert.

Dank schulde ich vielen Kollegen.

- Herrn Dr. D. Vennen, lange Jahre Mitarbeiter in unserer Abteilung, bin ich zu Dank verpflichtet für seine ungezählten Korrektur- und Ergänzungsvorschläge; er vor allem hat immer wieder darauf gedrungen, in den laufenden Text Beispiele aus der Praxis einzufügen.
- Herrn Professor Dr. R. Mausfeld danke ich für wiederholte Diskussionen über das Rasch-Modell, Herrn Dr. H. Stumpf für wiederholte Korrekturen des Kapitels über das Rasch-Modell.
- Ich war froh, daß ich Frau Dr. Pia Gottschalk und Frau Diplompsychologin Erika Haese die Transkription des ersten Manuskriptes anvertrauen konnte: Sie haben diese Aufgabe mit Sorgfalt gelöst.

Bei Frau Dr. Marion Bertgen-Giesen und Frau Diplompsychologin Mechthild Weidmann bedanke ich mich für ihre Textkorrekturen und ihre umfangreichen Literatur-Recherchen.

- Lebhaft danke ich auch Frau PD Dr. Cornelia von Hagen und ihrem Gatten, Herrn Diplompsychologen H. von Hagen, für ihre Informationsbeiträge und für mehrmalige Textkorrekturen.

Herr Dr. W. Jochmann von der Kienbaum Personalberatung hat zu jedem Abschnitt über „Bewerberbeurteilung“ in Kapitel 23 Anregungen gegeben, insbesondere zu den Themen „Rückmeldung, Anforderungsfacetten, Kandidatenklassifizierung, Einzel-Assessment, Selbstpräsentation“.

Bei der Vorbereitung der zweiten Auflage haben mich wiederum viele Kolleginnen und Kollegen unterstützt.

- Frau Dr. Ingrid Tismer-Puschner hat für viele Kapitel Korrekturen vorgeschlagen. Vor allem hat sie - dank einer vorzüglichen Kenntnis der Fachliteratur - ungezählte Ergänzungen und Erweiterungen angeregt, sie hat selber Texte beschafft und wichtige Stellen markiert: Sie hat mehr getan, als ihre Zeit ihr „eigentlich“ erlaubt hätte.
- Bei Herrn Dr. E. Fay vom Institut für Test- und Begabungsforschung in Bad Godesberg bedanke ich mich für unermüdliche Zusammenarbeit: Er hat Widersprüche in meiner Darstellung aufgedeckt, er hat Korrekturen vorgeschlagen, er hat Anregungen gegeben, er hat auf Literatur verwiesen - ungezählte Sätze und Abschnitte tragen seine Handschrift.
- Herr PD Dr. J. Funke hat mit mir die Problematik diskutiert, die sich ergibt bei Anwendung der klassischen Testtheorie auf Daten, die der Verhaltensbeobachtung oder einem Gespräch entstammen. Die entsprechenden Kapitel über Beobachtung und Interview hat er gegengelesen, für unzulängliche Angaben und Formulierungen hat er Korrekturen vorgeschlagen.
- Frau Diplompsychologin Eftychia Sidiropoulou hat es übernommen, für das Kapitel 18, „Computergestützte Diagnostik“, die Literatur zu suchen, sehr viele Beiträge zu sichten und die Ergebnisse in einer wohlgegliederten und umfangreichen Übersicht zu präsentieren. Darüber hinaus hat sie viele andere Kapitel kritisch gelesen und korrigiert.
- Herr Diplompsychologe Th. Fuchs hat in Kapitel 20 ein Beispiel für einen speziellen Fall von Begutachtung gestaltet; das Material beruht auf einem echten Fall aus seiner Praxis, doch hat er die personenbezogenen Daten so verändert, daß sich die Probandin nicht identifizieren läßt.
- Frau Diplompsychologin Andrea Obeldobel hat Literatur recherchiert und zusammengefaßt, sie hat so gut wie alle Kapitel gelesen und Fehler identifiziert: Sie mußte das Buch am gründlichsten kennen. Besonderer Dank gebührt ihr deswegen, weil sie diese Arbeit parallel zur Erstellung ihrer Diplomarbeit geleistet hat.
- Frau Katja Waligora hat, wie Frau Obeldobel, unermüdlich geholfen, Literatur zu beschaffen, sie war immer bereit, die gleichen Kapitel wieder

- und wieder zu lesen und die Fehler zu finden, die ich -bei jeder Korrektur
- wieder und wieder miteingefügt hatte.
- Mein Sohn Bernhard hat mir gezeigt, wie ich die Texte auf dem PC gestalten kann. Er hat mich in WordPerfect 6.1 für Windows eingeführt und Darstellungsprobleme gelöst, vor denen sein Vater kapitulierte.
 - Meine Frau Marlene hat die Entstehung des Werkes mit Geduld und Ironie begleitet. So hat sie mir geraten, Pausen einzulegen, um das Buch fertigzustellen. Außerdem hat sie immer wieder das weitere Vorgehen mitgeplant
- auf Kosten ihrer freien Zeit.

Nennen möchte ich auch die Mitarbeiter des Verlages für Psychologie. Herr Dr. H. Lundberg hat die Abfassung des Lehrbuches angeregt. Mit Herrn B. Otto hat er die ersten Konzeptionen gesichtet und mich zur Weiterarbeit ermutigt. Er hat das Werk dann seinem Nachfolger, Herrn Dr. M. Vogtmeier, empfohlen. Jedem von ihnen danke ich herzlich. Wenn das Buch einen Weg zu „seinem Leser“ findet, ist das auch ihr Verdienst.

Bei Frau Susanne Schurr und bei Herrn Hans-Joachim Bernert bedanke ich mich sehr herzlich: Die „Drucklegung“ und die Korrektur haben sie vorbildlich betreut. Sie haben nicht nur „Aufträge“ ausgeführt, sondern den Text und sein „Gesicht“ mitgestaltet. Wenn die Diagnostik „ansehnlich“ auftritt, dann auch, weil beide Mitgestalter ihr mit viel Geduld ein Ansehen verliehen haben.

Inhaltsverzeichnis

Vorwort		V
Inhaltsverzeichnis		IX
Teil I Vorfragen:		
	Gegenstandsbereich psychologischer Diagnostik und psychologischer Intervention	1
1. Kapitel: Zur Bestimmung		
	von Diagnostik und Intervention	3
1.1	Diagnostik und Intervention: Abgrenzungen (Definitionen)	3
1.2	Diagnostik und Intervention: Zur Entstehungsgeschichte	6
1.3	Finale, soziale, ethisch-juristische Struktur von Diagnostik und Intervention	8
1.4	Konzept einer Normativen Diagnostik	9
1.5	Zusammenfassung zu Kapitel 1	11
1.6	Kontrollfragen zu Kapitel 1	11
2. Kapitel: Diagnostik und Intervention		
	<i>Unterschiedliche Modellvorstellungen</i>	13
2.1	Konzepte zeitstabiler Eigenschaften	13
2.2	Prozeßorientierte Konzeptionen	14
2.2.1	Biographisch orientierte Persönlichkeitsmodelle	14
2.2.2	Psychodynamische Theorien	15
2.2.3	Kriteriumsorientierte Leistungsmessung	15
2.2.4	Interaktionistische Persönlichkeitsmodelle	16
2.3	Zusammenfassung zu Kapitel 2	17
2.4	Kontrollfragen zu Kapitel 2	17
3. Kapitel: Zur Darstellung psychologischer Diagnostik und Intervention		
	<i>Ansatz bei der Diagnostischen Situation</i>	19
	Zusammenfassung zu Kapitel 3	22

Teil II	Grundkenntnisse	23
4. Kapitel: Abriß der klassischen Testtheorie		31
4.1 Fragestellung, Testmerkmal, Test-Item		32
4.1.1 Konzeptualisierung von Fragestellung und Testmerkmal		32
4.1.2 Zuordnung von Testmerkmal und Test-Item		35
4.1.3 Wahl einer Konstruktionsstrategie		35
4.1.4 Bestimmung der Testart		36
4.1.5 Itemgenerierung und Itemgestaltung		38
4.2 Itemanalyse		40
4.2.1 Schwierigkeitsindex		41
4.2.1.1 Schwierigkeitsindex bei zweistufigen Antworten		41
4.2.1.2 Schwierigkeitsindex bei mehrstufigen Antworten		43
4.2.1.3 Erwünschte Schwierigkeitsindizes		46
4.2.1.4 Schwierigkeitsindex und andere Itemkennwerte		46
4.2.2 Trennschärfe		47
4.2.2.1 Berechnung der Trennschärfe		48
4.2.2.2 Teil-Ganz-Korrektur		51
4.2.2.3 Konvergente und diskriminante Trennschärfe		52
4.2.2.4 Trennschärfe und andere Itemkennwerte		54
4.2.3 Homogenität		54
4.2.3.1 Homogenität als Interkorrelation der Items		55
4.2.3.2 Homogenität im Sinne der Faktorenanalyse		56
4.2.3.3 Homogenität im Sinne einer Guttman-Skala		57
4.2.3.4 Homogenität im Sinne eines Rasch-Modells		59
4.2.4 Testrevision und Itemselektion		59
4.2.4.1 Inhaltliche Fragen der Itemselektion		59
4.2.4.2 Statistische Schritte der Itemselektion		60
4.2.4.3 Weitere Gesichtspunkte einer Itemselektion		65
4.3 Ermittlung der Test-Gütekriterien		66
4.3.1 Objektivität		66
4.3.1.1 Arten von Objektivität		67
4.3.1.2 Probleme der Objektivität		69
4.3.2 Reliabilität		70
4.3.2.1 Axiome der klassischen Testtheorie		70
4.3.2.2 Definition von Reliabilität		72
4.3.2.3 Veranschaulichung der Axiome und der Definition von Reliabilität		73
4.3.2.5 Modelle der Reliabilitätsberechnung		76
4.3.2.5 Test-Score und Vertrauensbereich		89
4.3.2.6 Kritische Differenz		92
4.3.3 Validität		93
4.3.3.1 Bestimmung (Definition) von Validität		94
4.3.3.2 Arten von Validität		95

4.3.3.3	Multitrait-Multimethod-Validierung: Paradigma einer Kombination von Validierungsarten	108
4.4	Normierung oder Eichung	111
4.4.1	Berechnung von Normen	111
4.4.1.1	Rohwerte	111
4.4.1.2	Transformierte Werte	112
4.4.1.3	Unterschiedliche Klassifikationen von Test-Scores und üblichen Normskalen	118
4.4.2	Probleme der Normierung	120
4.4.2.1	Wahl der Eichstichproben	120
4.4.2.2	Normalverteilte Merkmalsausprägung als Voraussetzung der Normierung	120
4.4.2.3	Normbezogene Klassifikation und Stichprobenabhängigkeit	121
4.4.2.4	Statistische Normen und kulturell-ethischer Kontext	122
4.5	Beitrag zu Diagnostik und Intervention	122
4.6	Kritik der klassischen Testtheorie	123
4.7	Zusammenfassung zu Kapitel 4	124
4.8	Kontrollfragen zu Kapitel 4	127

5. Kapitel: Hinweise

zur kriteriumsorientierten Leistungsmessung

5.1	Abgrenzungen: kriteriumsorientierte Leistungsmessung und kriteriumsorientierter Test	129
5.2	Konstruktion kriteriumsorientierter Testaufgaben	131
5.2.1	Operationale Definition	131
5.2.2	Aufspaltung der Aufgabe nach Zielen und Inhalten	132
5.2.3	Generative Regeln	132
5.3	Analyse kriteriumsorientierter Testaufgaben	135
5.3.1	Validität	136
5.3.2	Reliabilität	138
5.3.3	Objektivität	138
5.4	Schluß vom Testscore auf die Fähigkeit eines Probanden	140
5.4.1	Einstufige Entscheidung: Festlegung eines kritischen Punktwertes	141
5.4.2	Mehrstufige Entscheidung: Festlegung von Entscheidungsintervallen	141
5.4.3	Entscheidungen mit Vertrauensbereich	142
5.4.3.1	Bedeutung eines Vertrauensbereiches	142
5.4.3.2	Berechnung eines Vertrauensbereiches	143
5.5	Beitrag zu Diagnostik und Intervention	148
5.6	Zusammenfassung zu Kapitel 5	149
5.7	Kontrollfragen zu Kapitel 5	150

6. Kapitel: Der Grundgedanke des Rasch-Modells	151
6.1 Modellannahmen	151
6.2 Ausgangsgleichung	154
6.3 Schritte einer Rasch-Skalierung	157
6.3.1 Schritt I:	
Erstellung einer Matrix von Schwierigkeitsindizes (Matrix I)	157
6.3.2 Schritt II:	
Transformation von Matrix I in eine Logit-Matrix (Matrix II)	159
6.3.3 Schritt III:	
Schätzung der Personen- und Itemparameter aus Matrix II	161
6.3.4 Schritt IV - ein Modelltest: Reproduktion der Matrix I	164
6.3.5 Standardisierung	
der ermittelten Personen- und Itemparameter	169
6.4 Ergänzende Hinweise zur Endmatrix	171
6.4.1 Ermittlung von Vertrauensbereichen	171
6.4.2 Berechnung von Standardmeßfehlern	172
6.4.3 Iterative Berechnung von Modellparametern	173
6.4.4 Effektivere Algorithmen zur Parameterschätzung	173
6.5 Charakteristika einer Rasch-Skala	174
6.5.1 Homogenität	174
6.5.2 Lokale stochastische Unabhängigkeit der Items	176
6.5.3 Stichprobenunabhängigkeit von Skala und Items (spezifische Objektivität, Teilgruppenkonstanz)	177
6.5.4 Separierbarkeit von Item- und Personenparameter	178
6.6 Beitrag zu Diagnostik und Intervention	179
6.7 Kritische Anmerkungen	179
6.8 Zusammenfassung zu Kapitel 6	181
6.9 Kontrollfragen zu Kapitel 6	182
7. Kapitel: Verhaltensbeobachtung	183
7.1 Abgrenzungen (Definitionen)	183
7.2 Festlegung von Beobachtungseinheiten	187
7.3 Einteilung der Verhaltensbeobachtung	193
7.3.1 Systematische und unsystematische Beobachtung	193
7.3.2 Beobachtung von Verhaltensverlauf oder Verhaltenszustand	195
7.3.3 Beobachtung in natürlicher oder künstlicher Situation	195
7.3.4 Teilnehmende und nicht-teilnehmende Beobachtung	196
7.3.5 Erfassung von Zeit- oder Ereignisstichprobe	197
7.3.6 Verhaltensbeobachtung nach der Art ihrer Fixierung	197
7.4 Einfluß- und Verzerrungstendenzen	199
7.4.1 Allgemeine Fehler	199
7.4.2 Fehler speziell bei der Verhaltensbeobachtung	201

7.5	Beitrag zu Diagnostik und Intervention	203
7.6	Vor- und Nachteile der Verhaltensbeobachtung	206
7.7	Zu den Gütekriterien der Verhaltensbeobachtung	206
7.8	Zusammenfassung zu Kapitel 7	208
7.9	Kontrollfragen zu Kapitel 7	209

8. Kapitel: Gesprächsführung, Exploration, Interview, Anamneseerhebung

		211
8.1	Vorklärungen und Festlegungen	212
8.1.1	Abgrenzungen (Definitionsfragen)	212
8.1.2	Klassifikation von Gesprächen	214
8.1.2.1	Weiches, neutrales, hartes Gespräch	215
8.1.2.2	Standardisiertes, unstandardisiertes, halbstandardisiertes Gespräch	215
8.1.3	Explorative Fragetechniken	219
8.1.3.1	Klassifikation von Fragen	219
8.1.3.2	Formulierung von Fragen	223
8.2	Zur Praxis: Vorbereitung, Durchführung und Auswertung von Gesprächen	226
8.2.1	Vorbereitung von Gesprächen	227
8.2.2	Durchführung von Gesprächen	233
8.2.3	Zur Auswertung von Gesprächen in der Diagnostik	236
8.2.3.1	Wiedergabe des Originalgespräches	237
8.2.3.2	Zusammenfassung eines Gespräches	237
8.2.4	Beitrag zu Diagnostik und Intervention	243
8.2.5	Fehlertendenzen	247
8.3	Zu den Gütekriterien explorativer Daten	247
8.3.1	Explorative Daten und Objektivität	249
8.3.2	Explorative Daten und Reliabilität	250
8.3.3	Explorative Daten und Validität	253
8.4	Zusammenfassung zu Kapitel 8	256
8.5	Kontrollfragen zu Kapitel 8	257

Teil III Spezielle Einzelverfahren

9. Kapitel: Leistungstests	263
9.1 Allgemeine Charakteristika von Leistungstests	263
9.1.1 Definitionen und Abgrenzungen	263
9.1.2 Klassifikation von Leistungstests	267
9.1.3 Beitrag zu Diagnostik und Intervention	270
9.1.4 Aufgabenfelder für Leistungstests	273
9.1.5 Resümee zu Kapitel 9.1	274

9.2	Analyse und Vergleich von Testdaten:	
	Profilanalyse und Profilvergleich	275
9.2.1	Profilanalyse	275
9.2.2	Profilvergleich	282
9.2.3	Resümee zu Kapitel 9.2	288
9.3	Kriterien und Beispiel einer Testbewertung	288
9.3.1	Kriterien einer Testbewertung	288
9.3.2	Beispiel für eine Test-Bewertung (Fay, 1993)	291
9.3.3	Resümee zu Kapitel 9.3	296
9.4	Kontrollfragen zu Kapitel 9	296

10. Kapitel: Persönlichkeitstests, Fragebogen,

Persönlichkeitsinventare 297

10.1	Abgrenzungen: Eigenart des Fragebogens	297
10.1.1	Gemeinsamkeiten mit und Unterschiede zu Leistungstests	298
10.1.2	Konstruktion von Fragebogen	299
10.1.3	Sprachregeln zur Konstruktion von Fragebogen	300
10.2	Zur Beantwortung von Fragebogen	301
10.2.1	Kompetenz zur Selbstbeschreibung	301
10.2.2	Bereitschaft zur Selbstbeschreibung: Antworttendenzen	303
10.2.3	Relation von Selbstbeschreibung und Verhalten	306
10.3	Klassifikation von Fragebogen	307
10.4	Vorzüge und Nachteile von Fragebogen	309
10.4.1	Nachteile, Grenzen, Probleme	309
10.4.2	Chancen, Vorteile, Möglichkeiten	312
10.5	Beitrag zu Diagnostik und Intervention	313
10.6	Zusammenfassung zu Kapitel 10	315
10.7	Kontrollfragen zu Kapitel 10	316

11. Kapitel: Persönlichkeitsentfaltungsverfahren

oder projektive Verfahren 317

11.1	Abgrenzung des Konzeptes der Projektion	317
11.2	Klassifikation projektiver Verfahren	319
11.3	Probleme projektiver Verfahren	320
11.4	Beitrag projektiver Verfahren zur Diagnostik	321
11.5	Darstellung von drei Klassen projektiver Verfahren	322
11.5.1	Formdeuteverfahren	322
11.5.2	Verbal-thematische Verfahren	325
11.5.3	Zeichnerische und gestalterische Verfahren	330
11.6	Zusammenfassung zu Kapitel 11	332
11.7	Kontrollfragen zu Kapitel 11	333

Teil IV	Einzelaspekte integrativer Diagnostik	335
12. Kapitel:	Ethisch-juristische Determinanten von Diagnostik und Intervention	337
12.1	Zusammenfassung zu Kapitel 12	339
12.2	Kontrollfragen zu Kapitel 12	339
13. Kapitel:	Zwei diagnostische Grundaufgaben: <i>Klassifikation und Selektion</i>	341
13.1	Klassifikation	342
13.2	Selektion	345
13.3	Zusammenfassung zu Kapitel 13	349
13.4	Kontrollfragen zu Kapitel 13	349
14. Kapitel:	Zwei Wege der Entscheidungsfindung: <i>Statistische und Klinische Urteilsbildung</i>	351
14.1	Statistische Urteilsbildung	351
14.2	Klinische Urteilsbildung	352
14.3	Diagnostische Urteilsbildung und diagnostische Ziele	353
14.4	Vorrang des Statistischen Urteils?	354
14.5	Zusammenfassung zu Kapitel 14	355
14.6	Kontrollfragen zu Kapitel 14	355
15. Kapitel:	Drei Ansätze für Diagnostik und Intervention	357
15.1	Verhaltensperformanz oder Verhaltensdeskription	357
15.2	Synchrone oder diachrone Verhaltensbetrachtung	358
15.3	Verhaltensstatus oder Verhaltensprozeß	360
15.4	Zusammenfassung zu Kapitel 15	367
15.5	Kontrollfragen zu Kapitel 15	368
16. Kapitel:	Erfolgskontrolle	369
	Zusammenfassung zu Kapitel 16	371
	Kontrollfragen zu Kapitel 16	371
17. Kapitel:	Nutzenschätzung: <i>Entscheidungstheorie und Diagnostik oder Intervention</i>	373
17.1	Nutzenschätzung durch Zerlegung in Einzelkomponenten <i>Modell von Cronbach und Gleser</i>	374
17.2	Nutzungsschätzung durch Vergleich der Vorzüge verschiedener Methoden <i>„Multiattributive Nutzererschätzung“</i>	375

17.3	Nutzenschätzung durch Angaben in Geld	377
17.4	Nutzenschätzung durch Experten	377
17.5	Nutzenschätzung durch die Betroffenen	380
17.6	Zusammenfassung zu Kapitel 17	381
17.7	Kontrollfragen zu Kapitel 17	382
18.	Kapitel: Computerdiagnostik	
	<i>Eftychia Sidiropoulou</i>	383
18.1	Einsatz des Computers in der psychologischen Test-Diagnostik	383
18.2	Computersysteme	388
18.2.1	Computertests	389
18.2.1.1	Computer-Versionen von Papier-Bleistift-Tests	389
18.2.1.2	Computer-Simulationstests	390
18.2.1.3	Adaptive Tests	391
18.2.2	Computer-Testsysteme, Computer-Testgeräte	395
18.2.3	Computer-Interpretationssysteme	398
18.2.4	Computer-Expertensysteme	399
18.3	Einsatzfelder für eine computergestützte Diagnostik	401
18.4	Zusammenfassung zu Kapitel 18	402
18.5	Kontrollaufgaben zu Kapitel 18	403
Exkurs:	Zur Äquivalenz zwischen Papier-Bleistift-Tests und ihren Computer-Versionen	404

Teil V Integration:

Multimodale Diagnostik und Intervention . . 411

19.	Kapitel: Zum Verlauf integrativer Diagnostik	413
19.1	Verständigungsaufgabe - Struktur des diagnostischen Urteils .	413
19.2	Rahmenbedingungen der Psychologischen Situation	415
19.2.1	Allgemeine psychologische Determinanten	416
19.2.2	Spezielle psychologische Determinanten	417
19.3	Übersetzungsprobleme	418
19.3.1	Angemessene Fragestellung durch den Probanden	419
19.3.2	Angemessene Übersetzung durch den Diagnostiker	420
19.4	Verfahrensauswahl/Korrespondenzprobleme	421
19.4.1	Zuordnungsfrage: Korrespondenz von Problem und Verfahren	421
19.4.2	Anordnungsfrage: Korrespondenz von Problem und Sequenz der Verfahrensvorgabe	423
19.5	Integration der Ergebnisse	424
19.6	Vermittlung der Antwort an den Probanden	426

19.7	Intervention	426
19.8	Erfolgskontrolle: Evaluation und Supervision	427
19.9	Zusammenfassung zu Kapitel 19	428
19.10	Kontrollfragen zu Kapitel 19	429

20. Kapitel: Beispiel I für Integrative Diagnostik

Antrag auf Verlängerung einer Psychotherapie

Thomas Fuchs 431

20.1	Vorbemerkung	431
20.2	Einführung	432
20.3	Text des Verlängerungsberichts	432
	1. Zur Person	432
	2. Daten zur bisherigen und geplanten Behandlung	433
	3. Basisdaten zum Zeitpunkt des Behandlungsbeginns	433
	4. Diagnose mit ICD - Nummer (10. Revision)	435
	5. Angaben zur Genese und Aufrechterhaltung der Störungen/Symptomatik	435
	6. Behandlungsverlauf	436
	7. Therapieplanung	438
	8. Zusammenfassung	438

21. Kapitel: Beispiel II Integrativer Diagnostik

Psychologische Begutachtung 439

21.1	Abgrenzungen (Definitionen)	439
21.2	Psychologische Begutachtung im sozial-ethischen Kontext	441
21.3	Gutachten-Gliederung: <i>Überblick</i>	441
21.4	Gutachten-Gliederung: <i>Darstellung der einzelnen Abschnitte</i>	442
21.4.1	Erster Abschnitt des Gutachtens: <i>Übersicht</i>	442
21.4.2	Zweiter Abschnitt des Gutachtens: <i>Vorgeschichte</i>	444
21.4.3	Dritter Abschnitt des Gutachtens: <i>Untersuchungsbericht</i>	447
21.4.4	Vierter Abschnitt des Gutachtens: <i>Befund</i>	456
21.4.5	Fünfter Abschnitt des Gutachtens: <i>Stellungnahme</i>	466
21.5	Fehlertendenzen	471
21.6	Zusammenfassung zu Kapitel 21	474
21.7	Kontrollfragen zu Kapitel 21	475

22. Kapitel: Beispiel III Integrativer Diagnostik:

Beurteilung von Stellenbewerbern 477

22.1	Vorausgesetzte Situation und Aufgabenstellung	477
22.2	Erfassung der Stellenanforderungen	478
22.3	Vorauswahl der Bewerber	480
22.4	Beurteilung der ausgewählten Bewerber	482

22.4.1	Einige Voraussetzungen der Bewerberbeurteilung	482
22.4.2	Einzelschritte der Bewerberbeurteilung	483
22.4.3	Rückmeldung der Bewerberbeurteilung	484
22.4.4	Selbstpräsentation des Bewerbers vor dem Auftraggeber ...	486
22.4.5	Auswahlentscheidung	486
22.4.6	Nachbearbeitung	486
22.5	Unterrichtung abgelehnter Kandidaten	487
22.6	Evaluation, Erfolgskontrolle	487
22.7	Ethische Implikationen einer Bewerberbeurteilung	488
22.8	Zusammenfassung zu Kapitel 22	488
22.9	Kontrollfragen zu Kapitel 22	488
23. Kapitel: Beispiel IV Integrativer Diagnostik:		
	<i>Assessment-Center (AC)</i>	491
23.1	Abgrenzung (Definition)	491
23.2	Zeitliche Konzeption	494
23.3	Urteilsdimensionen	494
23.4	Übungen im Assessment-Center	495
23.5	Vorrang der Verhaltensbeobachtung	497
23.6	Ablaufbeispiel	498
23.7	Auswertung	499
23.8	Validität	500
23.9	Zusammenfassung zu Kapitel 23	501
23.10	Kontrollfragen zu Kapitel 23	501
Nachwort: Ratschlag an den Leser		503
Literatur		507
Personenregister		535
Sachregister		543

Teil I

Vorfragen: Gegenstandsbereich psychologischer Diagnostik und psychologischer Intervention

In Teil I versuchen wir, den Gegenstandsbereich von Diagnostik und Intervention zu umschreiben und zu gliedern. Wir besprechen

- die Bedeutung der Begriffe psychologische Diagnostik und psychologische Intervention (Kap. 1),
- unterschiedliche Modellvorstellungen von Diagnostik und Intervention (Kap. 2),
- die Gliederung der Stoffdarbietung in diesem Buche (Kap. 3).

1. Kapitel

Zur Bestimmung von Diagnostik und Intervention

Diagnostik und Intervention sind einander zugeordnet. Diagnostik soll zur Intervention führen, Intervention setzt Diagnostik voraus. Um diesen Zusammenhang zu verdeutlichen, erläutern wir

- Abgrenzungen von Diagnostik und Intervention (1.1),
- Beispiele zur Entstehungsgeschichte (1.2),
- finale, soziale, ethisch-juristische Struktur von Diagnostik und Intervention (1.3),
- Konzepte einer Normativen Diagnostik (1.4).

Das erste Kapitel schließt mit einer Zusammenfassung (1.5) und der Vorgabe einiger Kontrollfragen (1.6).

1.1 Diagnostik und Intervention: Abgrenzungen (Definitionen)

Das Wort Diagnostik geht zurück auf das griechische Verb ‚diagnoskein‘, das unterschiedliche Aspekte eines kognitiven Vorganges bezeichnet, vom Erkennen bis zum Beschließen. Das Verb bedeutet (1) genau kennenlernen, (2) entscheiden und (3) beschließen oder *sich* entscheiden (Kaegi, 1904, 184).

Diese drei Grundbedeutungen lassen vielfältige Assoziationen an Leistungen anklingen, die vom Psychologen als Diagnostiker erwartet werden: etwa, daß er menschliches Verhalten ‚gründlich kennenlerne‘, um bei Störungen zum Zwecke einer Abhilfe ‚Entscheidungen‘ oder gar ‚Beschlüsse‘ anzubieten.

Doch taugen solche etymologischen Ableitungen und ihre Assoziationen zu nicht mehr als zu Gedankenspielen. Denn die Begriffe ‚Diagnose‘ und ‚Diagnostik‘ haben eine Geschichte durchlaufen, während der sich ihre Bedeutung gewandelt hat.

Die Sachbedeutung hat sich verengt im Rahmen einer Fachsprache der Medizin: Diagnose und Diagnostik bezeichnen die Lehre und die Fertigkeit, Krankheiten zu erkennen und sie Ursachen oder Ursachensyndromen zuzuordnen.

In der Psychologie bezeichnet Diagnostik - befreit von dem Bezug zur Medizin - die Lehre von den Methoden und Verfahren der sachgemäßen Durchführung einer Diagnose. Eine ‚Diagnose‘ liefert Aussagen darüber, welche Sachverhalte (in der Vergangenheit) für ein Verhalten (in der Gegenwart) verantwortlich sind (Dorsch, 1994, 156; Schröder, 1976, 3-5). ‚Diagnostik‘ schließt heute auch Aussagen im Sinne einer Prognose ein.

Den Bedeutungshof mögen drei ‚Definitionen‘ veranschaulichen:

- *Psychologische Diagnostik ist das systematische Sammeln und Aufbereiten von Informationen mit dem Ziel, Entscheidungen und daraus resultierende Handlungen zu begründen, zu kontrollieren und zu optimieren. Solche Entscheidungen und Handlungen basieren auf einem komplexen Informationsverarbeitungsprozeß. In diesem Prozeß wird auf Regeln, Anleitungen, Algorithmen usw. zurückgegriffen. Man gewinnt damit psychologisch relevante Charakteristika von Merkmalsträgern und integriert gegebene Daten zu einem Urteil (Diagnose, Prognose). Als Merkmalsträger gelten Einzelpersonen, Personengruppen, Institutionen, Situationen, Gegenstände“ (Jäger R. S. & Petermann, 1995, 11).*
- *Psychodiagnostik ist eine Methodenlehre im Dienste der Angewandten Psychologie. Soweit Menschen die Merkmalsträger sind, besteht ihre Aufgabe darin, interindividuelle Unterschiede im Verhalten und Erleben sowie intraindividuelle Merkmale und Veränderungen einschließlich ihrer jeweils relevanten Bedingungen so zu erfassen, daß hinlänglich präzise Vorhersagen künftigen Verhaltens und Erlebens sowie deren evtl. Veränderungen in definierten Situationen möglich werden“ (Amelang & Zielinski, 1994, 3).*
- *Psychodiagnostik läßt sich definieren als ein Vorgehen, in dem menschliche Verhaltensdaten erhoben und auf der Grundlage von theoretisch-psychologischen Annahmen so interpretiert werden, daß sie eine Erklärung für vergangene und eine Vorraussage für zukünftige Verhaltensweisen erlauben. Außerdem sollen dem Diagnostizierten auf der Grundlage dieser Interpretationen geeignete Konsequenzen oder Behandlungen als Vorschlag unterbreitet oder sogar für ihn herbeigeführt werden“ (Ringelband & Birkhan, 1995, 796).*

Über einzelne Elemente der beiden ‚Definitionen‘ mögen Experten unterschiedlicher Meinung sein und sich deswegen auch streiten - insgesamt ergibt sich: Psychodiagnostik ist eine Methodologie, deren Aufgabe darin liegt, psychologisches Wissen und psychologische Techniken bereitzustellen, die dazu beitragen, (in Einzelfällen) praktische Probleme zu lösen (Westmeyer, 1993, 508).

Wie hebt sich von dieser Abgrenzung das Konzept der Intervention ab?

Das Wort **Intervention** leitet sich von dem lateinischen Verb *intervenire* ab, das soviel bedeutet wie: (1) in die Quere kommen, dazwischentreten, (2) unterbrechen, stören, hindern (Blase & Reeb, 1909, 434).

Im Angelsächsischen enthält das entsprechende Wort ‚intervention‘ eine ähnliche Bedeutung: „interferring or becoming involved, e. g. to prevent something happening“ (Homby, 1989, 658).

In beiden Fällen ergibt sich eine Grundbedeutung, die besagt, daß es um einen Eingriff geht, der einen Prozeßverlauf ändern und (dabei gegebenenfalls) Störungen beseitigen soll.

Dieser Grundbedeutung kommt es sehr nahe, wenn Intervention verstanden wird als ein „psychologisches Eingreifen, um die Entstehung oder das Andauern psychischer Störungen zu verhindern und diese letztlich abzubauen“ (Humboldt-Psychologie-Lexikon, 1990, 173). Damit wird Intervention zwar noch nicht gleichgesetzt mit Psychotherapie, aber doch in ihr Umfeld plaziert.

Muß indes das Konzept so eng gefaßt werden? Läßt sich das Konzept nicht auch weiter fassen, nämlich so, daß auch andere Maßnahmen als interventiv betrachtet werden: Maßnahmen, die einen psychischen Zustand *ändern* sollen?

Der Oberbegriff bezeichnet dann eine *Verhaltensänderung*. Das Konzept der Intervention nähert sich dem der Verhaltensmodifikation (Kaminski, 1970), es umfaßt Änderungswissen ebenso wie die Vertrautheit mit Änderungstechniken.

Intervention bezeichnet in der erweiterten Fassung ein psychologisches Handeln, das

- eine *Verhaltenänderung* anzielt,
- diese Veränderung systematisch *kontrolliert* und
- zur Herstellung oder Verbesserung des seelischen *Wohlbefindens* führt.

Amelang und Zielinski definieren Intervention wie folgt (1994, 263):

- *Interventionen sind „Maßnahmen, die aus den verschiedensten Gründen eingeleitet werden. Sie setzen an diagnostischen Feststellungen an, mit dem Ziel, Veränderungen auf organisatorischer oder individueller Ebene herbeizuführen. Im angloamerikanischen Raum ist dafür der Terminus ‚treatment‘, also Behandlung, gebräuchlich. Die intendierten Effekte sind erwartungsgemäß dann besonders positiv, wenn die Passung zwischen Diagnose und Intervention in optimaler Weise ausfällt.“*

Kasten 1-1 zählt drei Beispiele für Intervention auf.

Kasten 1-1 :**Beispiele für Intervention aus unterschiedlichen psychologischen Disziplinen**

- In der **Klinischen Psychologie** kann eine Beratung hinführen zu einer Therapie oder zu Schritten einer Gesundheitsprävention.
- In der **Arbeits- und Organisationspsychologie** kann ein *Trainingsprogramm* dazu beitragen, innerhalb eines Betriebes die Effizienz eines Einzelnen oder eines Teams zu steigern.
- In der **Werbepsychologie** kann eine Befragung dazu dienen, für ein bestimmtes Produkt ein Image zu *entwickeln*, das potentielle Käufer anziehen soll.

Resümee: Diagnostik und Intervention lassen sich verstehen als Abschnitte desselben psychologischen Prozesses, eines Ablaufs, in dem Intervention aus der Diagnostik hervorgeht. Diagnostik bezeichnet eher den erkundenden, Intervention eher den modifikatorischen Abschnitt dieser einheitlichen Handlungssequenz.

1.2 Diagnostik und Intervention: Zur Entstehungsgeschichte

Die Lehre von Diagnostik und Intervention ist weder ein Kind allein der Praxis noch allein der Theorie, sondern ein ‚Mischling‘ aus beidem, aber sie zielt darauf ab, Kenntnisse verschiedener Teildisziplinen der Psychologie für die Praxis des Lebens nutzbar zu machen.

Wir bringen einige *Hinweise* zur Entstehungs-Geschichte.

Außerhalb der Psychologie wurden Dienstleistungen umschrieben, welche die Psychologie erbringen könnte. *Innerhalb* der Psychologie wurden Theorien und Modelle entwickelt, die dazu anregten, theoretische Vorstellungen in konkreten Anwendungen auf die Probe zu stellen (Amelang & Zielinski, 1994, 3-6; Jäger, R. S. & Petermann, 1995, 17-48; Perrez & Baumann, 1991, 28; Thomae, 1977, 203-277; Wottawa & Hossiep, 1987, 5).

In jedem Beispiel verschränken sich Diagnostik und Intervention; ausdrücklich benannt sei die Verbindung nur in zwei Fällen.

Außerhalb der Psychologie wurden Dienstleistungen angefordert. Beispielsweise kamen Anfragen aus dem Gerichtssaal, aus dem pädagogischen Feld, aus den Personalbüros der Wirtschaft. Psychologen versuchten, solchen Erwartungen und Anforderungen zu entsprechen. Große Namen sind zu nennen:

- Als einer der ersten Psychologen stellte William Stern sein Fachwissen zur Verfügung, um die Aussagefähigkeit und Aussagegierlichkeit von Gerichtszeugen zu prüfen (Stern, 1904, 1926). *Anwendungsbeispiel:* Ein Psychologe habe die ‚Zurechnungsfähigkeit‘ eines Angeklagten festgestellt (Diagnostik). Dieses Urteil des Psychologen kann die Entscheidung der Richter beeinflussen und in die Urteilsbegründung eingehen (Intervention).

Im Auftrag des französischen Unterrichtsministeriums haben Binet und Simon den ‚Stufenleitertest der Intelligenz‘ („l'échelle metrique de l'intelligence“) entworfen und erprobt, um lernbehinderte Kinder zuverlässig von normalbegabten zu unterscheiden (Binet & Simon, 1905).

- Im Auftrag der amerikanischen Armee haben Woodworth und seine Mitarbeiter den ‚Persönlichen Datenbogen‘ (Personal Data Sheet) entwickelt. Unter den Männern, die sich 1917 um Aufnahme in das amerikanische Expeditionskorps für Europa bewarben, sollten der Datenbogen ungeeignete (neurotische) Anwärter identifizieren (Woodworth, 1919).
- Im Auftrag der Kultusminister der Länder der Bundesrepublik Deutschland hat ein Psychologen-Team den ‚Test für medizinische Studiengänge‘ (TMS) konzipiert und erprobt, damit zusätzlich zur Abiturnote ein Kriterium bei der Vergabe medizinischer Studienplätze herangezogen werden kann (Fay, 1982; Trost & Mitarbeiter, 1995).

Innerhalb der Psychologie wurden theoretische Annahmen und Modelle entwickelt, deren Anwendung in der Praxis sich anbot:

- Psychologen, die Theorien über die Intelligenz entwickelten, haben Anwendungsmodelle entworfen und in Meßverfahren erprobt. Diesen Bemühungen entsprangen viele Intelligenztests (Guilford & Hoepfner, 197, 1; Horn, 1983; Spearman, 1938; Thurstone, 1938).
- Aus persönlichkeits-theoretischen Ansätzen hat eine Vielzahl von Autoren eine Vielzahl von Fragebogen entworfen, die in der Praxis hilfreiche Dienste leisten (Beckmann, Brähler & Richter, 1990; Cattell et al., 1970; Eysenck, 1953; Guilford, 1959).
- Aus lerntheoretischen Konzepten, vor allem dem Konzept des operanten Konditionierens, hat Skinner das Prinzip der „Unterrichtsmaschinen“, der programmierten Bücher, auch des programmierten Unterrichts entwickelt (Skinner, 1948, 1953).
- Interaktionistisch orientierte Theoretiker haben sogenannte ‚Situations-Reaktions-Inventare‘ entwickelt, um Verhalten in seinen situativen Facetten zu erfassen (Petermann F. & U., 1980). *Anwendungsbeispiel:* Bei einem Jungen wird ermittelt, in welchen Situationen er auffällig aggressiv reagiert (Diagnostik). Aufgrund des Testwertes wird dem Probanden zu einem Verhaltenstraining geraten, das seine Aggressivität mindern soll (Intervention).

Resümee: Die Hinweise aus der Entstehungs-Geschichte sollten veranschaulichen, daß Diagnostik und Intervention weder allein aus der psychologischen Praxis noch allein aus der psychologischen Theorie hervorgegangen sind. Ihre Charakteristik besteht darin, verschiedene Teildisziplinen der Psychologie zu nutzen, um konkrete Lebenfragen lösen zu helfen.

1.3 Finale, soziale, ethisch-juristische Struktur von Diagnostik und Intervention

Eine diagnostisch-interventive Handlungssequenz schließt Momente ein, die finaler, sozialer, ethisch-juristischer Natur sind. Diese Qualitäten werden der Diagnostik und der Intervention nicht erst im nachhinein attribuiert.

Finale Struktur: Wenn Diagnostik ihre Hilfe anbietet, um Lebensfragen zu zu klären, und Intervention ein Programm skizziert, um Probleme zu bewältigen, dann sind finale Strukturen und Entscheidungsmomente mitenthalten. Ermittelt wird ja nicht nur, was ‚geschehen‘ ist und wie die Bedingungen des ‚Geschehens‘ aussehen (Kausal- und Bedingungsanalyse), bestimmt wird auch ein Ziel, auf das hin etwas geschehen soll (Handlungsvorschläge) (Jäger, R. S., 1985, 227; 1986, 13; Wottawa & Hossiep, 1987, 1, 18).

Woher aber kommt ‚Finalität‘ in die Diagnostik? Wenn sich Diagnostik begründet als Teilgebiet der Psychologie, dann muß die ‚Grundlage‘ ihrerseits ‚finale‘ Aussagen enthalten. Dazu nur zwei Hinweise:

- Motivationspsychologie schließt Zielorientierung ein. Motivationale Prozesse haben Zielcharakter. Wer also in der Diagnostik Motivationsprozesse aufdeckt, bringt auch Ziele zur Sprache.
- Persönlichkeitspsychologen legen finale Definitionen vor, ältere ebenso wie jüngere. Ältere, etwa Stern und Allport, bestimmen die ‚Person‘ als ‚Zielursache‘ (causa finalis). Jüngere, etwa Rotter und Mischel definieren die ‚Persönlichkeit‘ durch ‚Erwartungen‘, die sich auf zukünftige Verstärkungen richten.

In sich selber schließt der diagnostisch-interventive Prozeß finale Momente ein: Ein Klient oder Proband trage ein Anliegen vor, er suche die Hilfe des Psychologen, um die ‚richtige‘ Berufswahl zu treffen oder um eine ‚Verhaltensstörung‘ zu korrigieren. In solchen Fällen wird ‚etwas‘, was noch nicht existiert (‚Berufsbild‘ oder ‚normales‘ Verhalten) so antizipiert, daß es nur im antizipierenden Subjekt existiert und als intentionales Geschehen gegenwärtiges Verhalten auf erwünschte zukünftige ‚Zustände‘ ausrichtet.

Soziale Struktur: Zu den finalen kommen soziale Elemente hinzu. Diagnostisch-interventive Aufgaben beginnen, wenn sich ein Proband einer Frage gegenüberstellt, für die er den Rat des Psychologen sucht.

Diese ‚Frage‘ wird formuliert, zuerst in der Sprache des Betroffenen, dann in der Sprache des Diagnostikers. Dieser sprachliche Vorgang setzt Probanden und Diagnostiker in einen sozialen Kontext. Richtiger: In diesem Vorgang manifestiert sich die diagnostische Frage als eingebettet in einen sozialen Kontext, sowohl von der Seite des Probanden wie von der des Psychologen her.

Damit enthüllen sich Diagnostik und Intervention schon in ihrem Ansatz als partnerschaftliche Aufgabe. Nicht erst der Psychologe ‚erhebt‘ die Interaktion

zu einem sozialen Phänomen, von Anfang an ist das diagnostisch-interventive Handeln in einen partnerschaftlichen Kontext gesetzt.

Ethisch-juristische Struktur: In dieser finalen und sozialen ‚Bauform‘ von Diagnostik und Intervention sind ethisch-juristische Imperative enthalten.

(Amelang & Zielinski, 1994, 19-20; Booth, 1995, 138-147; Hartmann, 1984; Haubl, 1984; Jäger, R. S., 1986, 41-63; Klein, 1982; Petermann, 1995, 147-154; Pulver, Lang & Schmid, 1978; Schmid, 1995, 121-129; Schmidt, 1982; Schmidtchen, 1975, 36-40; Westhoff & Kluck, 1991; Wotawa & Hossiep, 1987, (73-89).

Der Psychologe hat in Diagnostik und Intervention mit Selbst- und Fremdbestimmung von Individuen zu tun, demnach mit ihrer Freiheit und Personwürde. Denn immer wieder werden ihm Informationen über persönliche, ja intime ‚Gegebenheiten‘ anvertraut, über Sachverhalte somit, die der Sphäre von Selbstbestimmung zugehören.

Diese Sphäre wird auch berührt, wenn der Untersucher eine Intervention vorschlägt, die etwa eine Berufs- oder Partnerwahl betrifft. Wiederum muß er sich fragen, wie weit seine Vorschläge die ‚Selbstverfügung‘ des Probanden respektieren.

Die Selbstenthüllung des Probanden und die Kenntnisnahme des Untersuchers vollziehen sich in einem vorgegebenen ethisch-juristischen Kontext.

Resümee: Eine diagnostisch-interventive Handlungssequenz schließt aus sich selber finale, soziale, ethisch-juristische Momente ein, wird durch sie selbst gleichsam mitkonstituiert. Nicht erst von außen, etwa zufolge der Entscheidungen eines Psychologen, erweisen sich Diagnostik und Intervention einem Rahmen zugeordnet, der auch andere Imperative einbezieht als die der Psychologie allein.

1.4 Konzept einer Normativen Diagnostik

Die vorausgegangenen Überlegungen laufen auf die Aussage hinaus, daß Diagnostik und Intervention in einem bestimmten Sinne immer ‚normativ‘ sind. Warum? Immer wieder muß der Psychologe erkennen, was ‚gegeben‘ ist, und muß sagen, was ‚geschehen‘ soll. Ständig muß er ‚Vergleiche‘ anstellen: zwischen einem Ist-Zustand und einem Soll-Zustand (etwa einer Störung und einem Zustand des Wohlbefindens). Erwartet wird, daß der Psychologe über ein ‚Kriterium‘ verfügt, an dem er diese ‚Vergleiche‘ legitimiert.

Genau diese Funktion bezeichnet das Konzept der ‚Norm‘: ein Begriff, der seiner Wortbedeutung nach soviel besagt wie Winkelmaß, Richtschnur, Regel, Vorschrift. Der Psychologe, der Diagnostik und Intervention betreibt, ist an

vielfältige und vielschichtige ‚Regeln und Vorschriften‘, in diesem Sinne also an ‚Normen‘ gebunden.

Freilich bedeutet ‚normativ‘ hier etwas anderes als etwa bei Westmeyer (1972), der als ‚Normative Diagnostik‘ ein Vorgehen beschreibt, das völlig regel- und theoriegeleitet ist, weil

- sowohl ‚*diagnostische Frage*‘ und ‚*diagnostische Antwort*‘
 - als auch der *Zusammenhang* zwischen Frage und Antwort
- eindeutig quantifizierbar und in Wahrscheinlichkeiten angebar sind.

Damit wird der diagnostische Akt unter dem Aspekt der Informationsverarbeitung betrachtet. Bei einer solchen Sicht ist vollständige Transparenz ein hohes Ziel.

In Grenzen zu erreichen und zu erstreben scheint diese Transparenz dann, wenn es um Verhaltensabläufe geht, die funktionalen Zusammenhängen gleich- oder nahekomen. Als Beispiel diene die kognitive Anforderung, die ein bestimmter Beruf stellt; einem Studenten, der Ingenieur werden will, aber mathematisch unbegabt ist, läßt sich nachweisen, daß sein Berufswunsch aus ‚funktionalen Gründen‘ unrealistisch ist. - Zwar müssen auch in solchen Fällen soziale und ethische Aspekte berücksichtigt werden, aber sie bleiben ‚verhüllter‘.

Für Fälle dieser Art stehen allerdings die Allsätze und Verknüpfungsregeln noch nicht zur Verfügung, die erforderlich wären. „Umfassende Anforderungsanalysen liegen für die Mehrzahl der Fragestellungen nicht vor“ (Durchholz, 1981, 273). Eine Normative Diagnostik im Sinne Westmeyers erweist sich darum zur Zeit als unrealisierbar - aus praktischen Gründen (Wottawa & Hosiepe, 1987, 59-60). „Davon unberührt bleibt aber der begrüßenswerte Versuch, heuristische Varianten innerhalb einer ‚Logik der Diagnostik‘ zur Diskussion gestellt zu haben“ (Guthke, Böttcher & Sprung, 1990, 40).

Doch gibt es auch Fälle, in denen es prinzipielle Gründe sind, die gegen eine Anwendung Normativer Diagnostik sprechen. Gedacht ist an Verhaltenszusammenhänge, in denen die Selbstbestimmung einer Person betroffen ist. Als Beispiel diene eine Partnerwahl. Falls die Rede von ‚Selbstbestimmung‘ einen Sinn behalten soll, dann gewiß den: daß in solchen ‚Wahlakten‘ das Verhalten etwas anderes ist als der Einzelfall eines allgemeinen Gesetzes. Darum scheint in einem solchen Kontext Normative Diagnostik unangemessen zu sein - selbst wenn in solchen Wahlakten auch funktionale Zusammenhänge zu berücksichtigen sind.

Freilich kommen an dieser Stelle anthropologische Überlegungen und Überzeugungen mit ins Spiel; soziale und ethisch-juristische Aspekte gewinnen eine größere Bedeutung als in Fällen, in denen der Diagnostiker nur ‚funktionale Zusammenhänge‘ erfassen soll.

1.5 Zusammenfassung zu Kapitel 1

Der Wortbedeutung nach geht Diagnostik auf ein griechisches Verb zurück, das soviel besagt wie ‚kennenlernen, beschließen, entscheiden‘. - Der Sachbedeutung nach bezeichnet Diagnostik eine Anwendungsmethodologie, die Regeln angibt, wie psychologische Charakteristika von Personen zu erfassen und Bedingungen ihres Verhaltens zu ermitteln sind.

Der Wortbedeutung nach geht Intervention auf ein lateinisches Verb zurück, das soviel bedeutet wie ‚einen Eingriff vornehmen, um einen Handlungsverlauf zu ändern und Störungen zu beseitigen‘. - Der Sachbedeutung nach bezeichnet Intervention ein psychologisches Handeln, das eine Verhaltensänderung anzielt, die das seelische Wohlbefinden verbessern soll; die Änderung muß systematisch kontrollierbar sein.

Diagnostische Aussagen und interventive Maßnahmen haben praktische Bedeutung für den Betroffenen. Darum stehen sie immer in einem finalen, sozialen und juristisch-ethischen Kontext.

Entwickelt haben sich Diagnostik und Intervention aus Anregungen, die von außerhalb und innerhalb der Psychologie kamen. Von außerhalb wurden Dienstleistungen angefordert, etwa aus dem Gerichtssaal oder aus dem pädagogischen Feld. - Innerhalb der Psychologie wurden theoretische Annahmen und Modelle entwickelt, die eine Anwendung in der Praxis nahelegten.

1.6 Kontrollfragen zu Kapitel 1

- Umschreibung psychologischer Diagnostik.
- Umschreibung psychologischer Intervention.
- Unterschiede zwischen Diagnostik und Intervention.
- Beispiele zur Entstehungsgeschichte von Diagnostik und Intervention.
- Finale, soziale, juristisch-ethische Struktur von Diagnostik und Intervention.
- Unterschiedliche Bedeutungen des Konzeptes ‚Normative Diagnostik‘.

2. Kapitel

Diagnostik und Intervention

Unterschiedliche Modellvorstellungen

Diagnostik und Intervention bilden kein einheitliches System, sie gehen auf unterschiedliche Ansätze zurück.

„Psychologische Diagnostik entwickelte sich in unserem Jahrhundert zunächst als eine besonders hoffnungsvolle Teildisziplin der Psychologie. Heute steht sie in vielfältigen Auseinandersetzungen mit divergenten persönlichkeits- und verhaltenstheoretischen Positionen, beruht auf unterschiedlichen methodologischen Ansätzen und wird mit berufsethischen und gesellschaftspolitischen Problemen konfrontiert. Sie hat sich gleichermaßen gegenüber Erwartungen wie gegen globale Disqualifikationen zu wehren“, (Groffmann & Michel, 1982 a, VII; vgl. Guthke, Böttcher & Sprung, 1990, 23-29; Jäger R. S. & Petermann, 1995, 15-48, 77-117; Leichner 1979, 8-9).

Kapitel 2 skizziert zwei persönlichkeitspsychologische Ansätze, welche die Entwicklung des diagnostisch-interventiven Instrumentariums entscheidend geprägt haben:

- Persönlichkeitstheorien, die zeitstabile Eigenschaften annehmen (2.1),
- und Theorien, die eine Person als Prozeßgestalt deuten (2.2).

Das Kapitel schließt mit einer Zusammenfassung (2.3) und der Vorgabe einiger Kontrollfragen (2.4)

2.1 Konzepte zeitstabiler Eigenschaften

Die Diagnostik wurde von Theoretikern mitgeprägt, die annahmen, menschliches Verhalten entspringe sogenannten Eigenschaften (traits), die sich als relativ zeitstabil erweisen. Dem Verhalten wird zwar eine gewisse Variationsspanne zuerkannt. Vereinfacht gilt jedoch, daß sich eine Person gleichartig verhält über Situationen und über Zeiten hinweg. Zu ermitteln sind darum Eigenschaften, denen das konstante und konsistente Verhalten entspringt.

„Ein orthodoxer trait-Ansatz postuliert, daß Verhalten ausschließlich vom trait-Wert abhängig ist; trait und Verhalten stehen in monotoner Beziehung.

Situationen nehmen keinen modifizierenden Einfluß“ (Leichner, 1979, 29). -Allerdings ist zu ergänzen: „Der wechselseitige Einfluß von Eigenschaften und Situationen erzeugt vorübergehend innere Bedingungen, die als Zustände (states) bekannt sind“ (Eysenck & Eysenck, 1987, 35).

Diesem Ansatz wird eine bestimmte Art von Diagnostik zugeordnet - zusammengefaßt in den Aussagen der klassischen Testtheorie und realisiert in Verfahren, deren Konstruktion sich an ihr orientiert (Kap. 4, S. 31).

Das Instrumentarium der klassischen Testtheorie ist heftig kritisiert worden (Fischer, 1974, 16-145; Goldfried & Kent, 1976; Grubitzsch, 1991; Pawlik, 1976; Schaller & Schmidtke, 1983, 491-508).

Für bestimmte Fragestellungen bleibt die Annahme relativ zeitstabiler Eigenschaften jedoch sinnvoll: Beispiele sind Eignungsprüfungen, Beratungen zum Verlauf der Bildungskarriere, Fragen der Forensischen Psychologie.

Alleinbestimmend war dieser Ansatz nicht.

2.2 Prozeßorientierte Konzeptionen

Persönlichkeitstheorien, die menschliches Verhalten vor allem als Prozeßgestalt interpretieren, unterscheiden sich vielfältig. Auf vier Spielarten sei verwiesen, auf

- biographisch orientierte Modelle (2.2.1),
- psychodynamische Theorien (2.2.2),
- kriteriumsorientierte Leistungsmessung (2.2.3),
- interaktionistische Ansätze (2.2.4).

2.2.1 Biographisch orientierte Persönlichkeitsmodelle

Die Benennung ‚biographisch orientierte Persönlichkeitsmodelle‘ soll eine Gruppe von Theoretikern bezeichnen, die menschliches Verhalten zu verstehen suchen von seiner biographischen Genese her. Gedacht ist an Vertreter wie Bühler, Ch. (1933, 1969) oder Freud (1940), Fuchs, W. (1982), Kelly (1955) oder Murray (1938), Stern (1921, 1923) oder Thomae (1968) (vgl. Jüttemann & Thomae, 1987).

Biographische Forschung hat eine Affinität zu explorativen oder explorationsähnlichen Methoden, also zu Gespräch, Befragung, Analyse von Selbstbeschreibungen (Tagebüchern, Briefen, persönlichen Dokumenten).

Diese Verfahren sind ebenso nachdrücklich abgelehnt (Eysenck, 1967) wie entschieden verteidigt worden (Kruse, 1987; Lehr, 1964; Mischel, 1993; Thomae, 1968). Je ‚existentieller‘ jedoch das diagnostische Problem ist, das ein

Proband zur Sprache bringt, desto höher dürfte die Bedeutung sein, die dem ‚Gespräch‘ zufällt (Kap. 8, S. 211).

2.2.2 Psychodynamische Theorien

Auch psychodynamische Theorien versuchen, Verhalten aus der Biographie eines Menschen zu begreifen. Darüber hinaus betonen sie aber, Verhalten werde gesteuert von unbewußten Spannungen und ihren Entladungen. Um Verhalten zu begreifen, müsse der Psychologe demnach die unbewußten Prozesse ermitteln.

Diese Prozesse lassen sich jedoch nicht so erfassen, wie man Eigenschaften mißt. Zwar ‚äußern‘ sie sich in Verhaltensweisen wie Versprechen, Träumen oder neurotischen Symptomen. Aber sie bis zu ihrem unbewußten Ursprung zu verfolgen erfordert voraussetzungsvolle Interpretationsschritte - die mitzugehen nicht jeder Psychologe bereit ist.

Psychodynamische Theorien werden von unterschiedlichen Autoren vertreten. Erwähnt seien nur zwei Gruppen:

1. die drei *Klassiker Freud, Adler, Jung* und die verschiedenen tiefenpsychologischen Schulen, die sich gebildet haben (vgl. etwa Grawe, Donati & Bemauer, 1994; Kriz, 1991; Wyss, 1966);
2. die *Humanistischen Psychologen*, etwa Rogers, Maslow, Fromm und die unterschiedlichen Gruppen ihrer Anhänger (vgl. etwa Kriz, 1991; Quitmann, 1991; Volker, 1980).

Zur Erfassung unbewußter Dynamismen wurde eine eigene Klasse diagnostischer Instrumente entwickelt: die sogenannten projektiven Methoden, etwa sogenannte ‚thematische‘ Verfahren. Heftiger noch als um die klassischen Tests ist um die projektiven Verfahren gestritten worden (Axhausen, 1989; Hörmann, 1982; Lechner, 1983).

Gegen alle Einwände bleibt festzuhalten: Bei bestimmten Fragestellungen, vor allem klinischer Natur, können sie hilfreiche Suchdienste übernehmen, sie können Heurismen für das weitere Vorgehen bereitstellen (Kap. 11, S. 317).

2.2.3 Kriteriumsorientierte Leistungsmessung

Als Gegenpart zu den Theorien, die Verhalten auf zeitstabile Merkmale oder auf unbewußte Prozesse zurückführen, wurde ein diagnostisches Modell konzipiert, das kriteriumsorientierte Leistungsmessung‘ anzielt. Wie bei psychodynamischen Ansätzen soll der Prozeß des Verhaltens erfaßt werden, diese Erfassung soll sich aber orientieren an einem wohldefinierten Kriterium, etwa einem pädagogischen oder einem therapeutischen Ziel (Fricke, 1974; Glaser, 1973; Klauer, 1987).

Diagnostisches Instrument ist darum eine exakte Beobachtung und Beschreibung menschlichen Verhaltens. Zwar werden auch Tests eingesetzt - diese Tests sind jedoch nicht konzipiert nach der klassischen Testtheorie, sondern orientiert an einem Kriterium: einem Ziel, zu dem eine Therapie oder ein pädagogischer Prozeß hinführen soll. Der kriterienorientierte Test repräsentiert eine Stichprobe des Zielverhaltens; der Testwert dient als Indikator dafür, wie weit sich ein Proband dem Ziel genähert hat (Kap. 5, S. 129).

2.2.4 Interaktionistische Persönlichkeitsmodelle

Interaktionistische Persönlichkeitsmodelle stellen das Verhalten dar als Resultante von Person und Situation, in diesem Sinne als Interaktion. Zu erfassen sind demnach ‚gleiche‘ Verhaltensweisen, die in unterschiedlichen Situationen auftreten, oder ‚gleiche‘ Situationen, die unterschiedliche Verhaltensmuster hervorrufen. In diesem Modell verbinden sich lerntheoretische und kognitionspsychologische Ansätze - es handelt sich um sozial-kognitive Lerntheorien, beispielsweise von Rotter und Hochreich (1979) oder Bandura (1977) und Mischel (1993).

Als ein diagnostisches Instrument, das diesem Ansatz affin ist, wurde das sogenannte ‚Situations-Reaktions-Inventar‘ entwickelt, ein Fragebogen, der ermitteln soll, wie sich bestimmte Formen gleichen Verhaltens in verschiedenen Situationen äußert (Noack & Petermann, 1995; Petermann, F. & U., 1978, 52-53; Petermann, F. & U., 1987).

Resümee: Schon diese zwei großen Gruppen von Ansätzen - erstens die Konzepte zeitstabiler Eigenschaften, zweitens die prozeßorientierten Persönlichkeitskonzeptionen - lassen sich nicht zu einem einheitlichen System von Diagnostik und Intervention integrieren. Erst recht gelingt die Integration dann nicht, wenn andere Aspekte auch berücksichtigt werden, etwa wissenschafts- oder meßtheoretische Schulmeinungen.

Im einzelnen diagnostischen Schritt, in der einzelnen diagnostischen Situation wird einmal der eine Ansatz überwiegen (etwa die Anwendung von Tests), ein andermal der andere Ansatz (etwa der Einsatz projektiver Verfahren). Komplexe Fragestellungen nötigen den Diagnostiker meist dazu, verschiedenen Ansätzen zu folgen - eine Aufgabe, die ihm ständige ‚Systemüberschreitungen‘ abverlangt.

2.3 Zusammenfassung zu Kapitel 2

Psychologische Diagnostik und Intervention ergeben kein einheitliches System. Zwei große unterschiedliche persönlichkeits-theoretische Ansätze haben sie entscheidend geprägt:

- Einem Konzept, das von zeitstabilen Eigenschaften ausgeht, entstammt vor allem jenes Instrumentarium, das sich an der klassischen Testtheorie orientiert.
- Konzeptionen, welche die Person von ihren Verhaltensabläufen her deuten, begünstigten die Entwicklung und die Verwendung höchst unterschiedlicher Methoden. Einige Theoretiker favorisierten die Exploration, andere die Verhaltensbeobachtung, wieder andere die kriteriumsorientierten Tests, die Situations-Reaktions-Inventare oder die projektiven Verfahren.

2.4 Kontrollfragen zu Kapitel 2

- Modellvorstellungen psychologischer Diagnostik und Intervention,
- Modelle zeitstabiler Eigenschaften und ihre Instrumente.
- Modelle prozeßhafter Merkmale und ihre Instrumente.
- Biographisch orientierte Modelle und ihre Instrumente.
- Psychodynamische Theorien und ihre Instrumente.
- Kriteriumsorientierte Leistungsmessung.
- Interaktionistische Ansätze und ihre Instrumente.

3. Kapitel

Zur Darstellung psychologischer Diagnostik und Intervention

Ansatz bei der Diagnostischen Situation

Da es kein einheitliches Konzept psychologischer Diagnostik und Intervention gibt, ist nicht zu vermuten daß ihre Darstellung einheitlich ausfällt.

Die Sichtung einiger Lehrbücher bestätigt diese Vermutung (Amelang & Zielinski, 1994; Jäger, R. S. & Petermann, 1995; Groffmann & Michel, 1982, 1983; Guthke, Böttcher & Sprung, 1990, 1991; Kubinger, 1995 b; Lechner, 1979; Wehner, 1981; Wottawa & Hossiep, 1987).

Erkennbar werden aber Grundlinien:

- Dargestellt wird die historisch-systematische Einbettung von Diagnostik und Intervention, etwa die Herkunft aus der Psychiatrie oder die Verwandtschaft mit der Persönlichkeitspsychologie, die Einbindung psychologischer Arbeit in Ethik und Recht.
- Ausführlich werden methodische Fragen abgehandelt, etwa die Grundlagen der Testtheorien, die Unterschiede der Datenklassen, die Aufgabe von Hypothese und Erklärung im diagnostischen Prozeß.
- Einzelne Verfahrensklassen werden breit geschildert, mit ihrer Geschichte, mit ihren diagnostischen Vorteilen und ihrer Problematik, etwa Leistungstests, Persönlichkeitsinventare oder Interviewtechniken.
- Zur Sprache kommen Prozeduren, die Synthesen verlangen, etwa Klassifikationsverfahren, Formen diagnostischer Urteilsbildung, Modelle der Gutachtererstellung.
- Schließlich werden Anwendungsfelder skizziert, etwa die Rolle von Diagnostik und Intervention in der Pädagogischen oder Klinischen Psychologie, die Aufgabe von Diagnostik und Intervention in der Forensischen Psychologie oder in der Arbeits- und Organisationspsychologie.

Einschränkung: Eine Auflistung wie diese verschleiert den Umfang und die Vielschichtigkeit, die Komplexität und die Disparatheit der diagnostischen und interventiven Thematiken, welche die Lehrbücher vortragen - die Auflistung vereinfacht, vielleicht simplifiziert sie auch. Die Vereinfachung soll einen Überblick ermöglichen und eine Gliederung erleichtern.

In den Lehrbüchern überwiegt die Darstellung einer ‚formalen‘ Diagnostik. Das soll besagen: Die Lehrbücher beschreiben detailliert, wie der Psychologe *generell* eine Fragestellung aufschlüsseln und eine Antwort suchen kann. Sie leiten ihn an, wie er *generell* Fragebogen bewerten oder Explorationstechniken anwenden soll. Seltener beschreiben sie, wie er einzelne Merkmalsklassen definieren kann und welche Einzelverfahren zu ihrer Erfassung vorliegen.

Eine Lehre, welche einzelne Merkmale und die Verfahren zu ihrer Erfassung bespricht, also eine ‚materiale‘ Lehre von Diagnostik und Intervention, liegt nur in Ansätzen vor:

- Behandelt werden ‚klassische‘ Merkmalgruppen wie Intelligenz, Konzentration und andere Leistungsfunktionen; vorgestellt werden Verfahren, die zu ihrer Erfassung entwickelt wurden, etwa Intelligenz- und andere Leistungstests.
- Eher sparsam besprochen werden Persönlichkeitsmerkmale wie Angst oder Aggression, Leistungsmotivation oder Extra/Introversion; Verfahren, die zu ihrer Erfassung dienen könnten, werden nur am Rande erwähnt.

Auch dieses Buch folgt dieser Linie: Insgesamt wird eine formale Lehre von Diagnostik und Intervention geboten.

Suche nach einer Gliederung: Die Lehrbücher bieten viele Anhaltspunkte für eine Gliederung. Um weitere Gesichtspunkte zu entdecken, betrachten wir eine ‚typische‘, somit vereinfachte diagnostische Untersuchung. Von der vereinfachten diagnostischen Situation her wollen wir die weitere Gliederung entwerfen.

Diagnostische Situation ist „ein Sammelname für eine Reihe von Untergruppen von Situationen, deren jeweilige Bezeichnung sich nach den eingesetzten Verfahren (z. B. Test - bzw. Interviewsituation) oder nach der Aufgabenstellung (z. B. Examenssituation, Eignungsuntersuchungs- oder experimentelle Situation) oder nach der besonderen Aktivität (z. B. Vortragssituation) richten kann“ (Spitznagel, 1982 a, 250).

In Kasten 3-1 sei eine diagnostische Untersuchung in einen schematischen Ablauf umgesetzt.

Kasten 3-1:

Diagnostische Untersuchung - vereinfacht und schematisiert

- | |
|--|
| <ul style="list-style-type: none"> - Eine Fragestellung wird von einem Probanden eingebracht, eine Lösung oder Beantwortung von einem Psychologen als Experten erwartet.
Die Fragestellung übersetzt der Psychologe in ein psychologisches Untersuchungsszenario. — Die entsprechenden psychologischen Untersuchungsverfahren müssen bestimmt werden, beispielsweise Tests, Fragebögen, Interviews. — Es folgt eine Phase der Untersuchung, zum Beispiel des Testens, der Verhaltensbeobachtung, der Anwendung apparativer Verfahren. — Die erhobenen Daten müssen ausgewertet, die Ergebnisse verglichen und interpretiert werden. — Zu entscheiden ist dann, ob die gewonnenen Informationen ausreichen, um die Ausgangsfrage zu beantworten, oder ob neue Informationen einzuholen sind. |
|--|

- Reicht die gesammelte Information aus, muß der Psychologe eine Antwort auf die zu Beginn gestellte Frage formulieren, beispielsweise eine Diagnose erstellen, eine Prognose geben, einen Entscheidungsvorschlag unterbreiten, interventive Maßnahmen empfehlen.
- Soweit möglich, muß er sich des Erfolgs vergewissern; er muß prüfen, ob seine diagnostisch-interventive Handlungssequenz in ‚Erfolg‘ mündete oder bei ‚Mißerfolg‘ endete: Er muß seine Ergebnisse evaluieren.

Gehen wir das Szenario in Kasten 3-1 durch: Was hat der Diagnostiker in der diagnostischen Situation zu leisten?

Beginnen wir am Ende:

1. *Auf die ‚Ausgangsfrage‘ eine ‚psychologische Antwort‘ zu geben, erfordert die Synthese unterschiedlicher Informationen. Der Psychologe muß unterschiedliche Verfahren anwenden und ihre Aussagen integrieren. Diese Leistung läßt sich darum zusammenfassen unter dem Titel der **Integration**.*
2. *Die Integration stützt sich auf die Ergebnisse einzelner Verfahren, etwa auf Test- und Fragebogenwerte oder auf Daten, die aus projektiven Verfahren stammen. Der Diagnostiker muß demnach über die **Kenntnis spezieller Einzelverfahren** verfügen,*
3. *Die Anwendung einzelner Verfahren setzt voraus, daß der Psychologe bestimmte **Grundkenntnisse** erworben hat: Beispielsweise erfordert der Einsatz von Tests die Kenntnis der Testtheorien, ein Interview verlangt die Beherrschung von Regeln der Gesprächsführung, eine Vorlage projektiver Verfahren schließt die Vertrautheit mit ihren Anwendungs- und Auswertungsregeln ein.*

Weitere Gliederung dieses Buches

An diesen drei Aspekten orientiert sich die weitere Gliederung des Buches. Wir referieren den Lehrstoff in drei Durchgängen:

- Unter dem Namen **diagnostischer Grundkenntnisse** werden Testtheorien, Regeln der Verhaltensbeobachtung und der Gesprächsführung vorgestellt (Teil II).
- Unter dem Titel **spezieller Einzelverfahren** werden *zum einen* Leistungs- und Persönlichkeitstests, *zum anderen* projektive Verfahren besprochen (Teil III).
- In einem weiteren Durchgang besprechen wir **Einzelfragen**, die in den anderen Teilen zwar erwähnt, aber nicht explizit behandelt werden. Genannt seien Aufgaben wie Klassifikation und Selektion, Statistische und Klinische Urteilsbildung, Erfolgskontrolle, Nutzenschätzung, edv-gestützte Diagnostik (Teil IV).
- Als Aufgabe einer **diagnostisch-interventiven Integration** werden Untersuchungsprozeß und Urteilsbildung beschrieben und Beispiele für diese Leistung angeführt, etwa der Verlauf eines Assessment-Centers oder die Erstellung eines psychologischen Gutachtens (Teil V).

Zusammenfassung zu Kapitel 3

Es wird die weitere Gliederung des Stoffes in diesem Buche vorgestellt. Der Lehrstoff wird in drei Durchgängen geboten:

- Vorgestellt werden ‚Grundkenntnisse von Diagnostik und Intervention‘; dazu zählen psychologische Testtheorien sowie Regeln der Verhaltensbeobachtung und der Gesprächsführung (Teil II).
- Als ‚spezielle Einzelverfahren‘ werden die Klassen der Leistungs- und Persönlichkeitstests sowie der projektiven Verfahren besprochen (Teil III).
- In einem eigenen Durchgang werden Einzelfragen besprochen, die in den anderen Teilen zwar vorkommen, aber nicht explizit behandelt werden (Teil IV).
- Schließlich werden Schritte und Beispiele ‚diagnostischer Integration‘ beschrieben (Teil V).

Teil II

Grundkenntnisse

Der Titel „Grundkenntnisse“ soll jene Wissensanteile bezeichnen, die jeder diagnostisch-interventive Schritt einschließt. Drei Bereiche seien dazu gezählt: Verhaltensbeobachtung, Gesprächsführung, Testtheorien.

Begründung der Stoffauswahl: Auf ‚*Gespräche*‘ ist jeder Anwender angewiesen - etwa wenn er klaren will, warum ein Klient ihn aufsucht. Aussagen, die aus Gesprächen stammen, lassen sich ergänzen und bereichern durch Aussagen, die auf ‚*Verhaltensbeobachtungen*‘ beruhen. Insofern gehören Kenntnisse über diese beiden Verfahren zum ‚Grundwissen‘ eines diagnostisch und interventiv tätigen Psychologen.

Gespräch und Beobachtung lassen sich in Wechselbeziehung zu den *Testtheorien* setzen. Wer Daten quantifizieren will, die er in einer Exploration oder bei Verhaltensbeobachtungen gewonnen hat, sieht sich auf Meß- oder Testtheorien verwiesen, setzt ihre Geltung also voraus.

Umgekehrt gilt: Wer Meß- oder Testtheorien auf Daten anwendet, setzt inhaltlich verbale Klassifikationen voraus - gewonnen beispielsweise bei Verhaltensbeobachtung oder in Interviews.

Weil aber die Behandlung von Gespräch und Verhaltensbeobachtung schon testtheoretische Fachbegriffe erfordert, etwa Objektivität oder Validität, sei vorgegangen wie folgt:

- Zunächst seien Testtheorien skizziert (Kap. 4-6),
- danach die Verhaltensbeobachtung behandelt
- schließlich die Gesprächsführung vorgestellt (Kap. 8),

Zwei Grundprobleme: Bei Beschäftigung mit Testtheorien, Verhaltensbeobachtungen und Gesprächen trifft der Diagnostiker ständig auf zwei Probleme:

- In der Psychologie sind die **Variablen** nicht vorgegeben, sondern müssen ‚*konstruiert*‘ werden. (In der Physik sind manche Größen vorgegeben, wenigstens für den ersten Blick, etwa Länge, Breite, Höhe.) In der Psychologie ist kein ‚Gegenstand an sich‘ gegeben, weder ‚Intelligenz‘ oder ‚Konzentration‘ noch ‚Stimmung‘ oder ‚Zufriedenheit‘. Der messende Psychologe muß seine Gegenstände immer wieder neu definieren - ‚abgrenzen‘.

Darum nötigt fast jedes der folgenden Kapitel zur Besprechung von Definitionsfragen, von Versuchen, ‚Variablen zu operationalisieren‘.

„Da es keine gültigen Kriterien darüber gibt, wann die Operationalisierung einer Variablen gelungen ist, ist eine genaue Beschreibung dieser Operationalisierung unerlässlich“ (Wittke, 1980, 29).

- Die neu definierten **Variablen** betreffen Prozesse, die sich **ändern** oder ändern können, **indem sie beobachtet oder gemessen werden**. Es ist das Problem der Reagibilität, der ‚betroffenen‘ Reaktion auf einen Meßvorgang. Der Gegenstand ‚Intelligenz‘, wird er gemessen, ändert sich, beispielsweise durch Lernen: Ein Schüler, dem zweimal derselbe Schulleistungstest vorgelegt wird, überträgt aus der ersten Testung Erfahrungen auf die zweite Testung. - In der diagnostischen Situation ergibt sich somit das Problem, daß sich bei derselben Person kaum je ein Verfahren beliebig wiederholen läßt (wie etwa in der Makrophysik eine Längenmessung).

Diese beiden Probleme werden die nun folgende Darstellung beeinflussen, zunächst die der Testtheorien.

Vorbemerkung zu den Testtheorien: In vielen diagnostischen Untersuchungen werden Tests verwandt: Verfahren, die Leistungen messen oder Persönlichkeitsmerkmale erfassen. Der Konstruktion solcher Verfahren liegen Regeln zugrunde, die in Testtheorien formalisiert sind.

Wir skizzieren drei Ansätze:

- die klassische Testtheorie (Kap. 4),
- die kriteriumsorientierte Leistungsmessung (Kap. 5),
- das probabilistische Modell von Rasch (Kap. 6).

Den drei Kapiteln seien drei Abschnitte vorausgeschickt:

- eine Umschreibung von Tests,
- eine Einteilung psychodiagnostischer Verfahren und
- ein Hinweis auf Messen als Voraussetzung einer Testkonstruktion.

Umschreibung von Tests

Das Wort ‚Test‘ hat vielerlei Bedeutungen, in der Umgangssprache wie in verschiedenen Fachsprachen. Darum sei festgelegt, was hier unter psychologischen Tests verstanden wird.

Da zur Zeit die meisten Tests noch nach den Regeln der klassischen Testtheorie konstruiert sein dürften, sei eine Umschreibung gewählt, die dem ‚klassischen‘ Ansatz zugeordnet ist, sie stammt von Guthke (1972, 69). Die Definition erlaubt es aber, Unterschiede zu den beiden anderen Theorien (kriteriumsorientiert, probabilistisch) zu markieren.

Ais Test soll gelten „ein Prüfverfahren, bei dem in standardisierten Situationen Verhaltensmerkmale (Verhaltensstichproben) von Personen erfaßt werden, die als Indikatoren für bestimmte Personeneigenschaften dienen sollen, und dessen Resultate eine Einordnung der Untersuchten in eine Klassifikation ermöglichen, die an einer Gruppe vergleichbarer Personen gewonnen wurde“.

In der Umschreibung treten vier Charakteristika hervor:

1. Ein psychologischer Test ist ein **Prüfverfahren**: Er dient praktischen oder wissenschaftlichen Unterscheidungszwecken. Er soll es erlauben, Personen nach bestimmten Merkmalsausprägungen zu unterscheiden, so wie eine Klassenarbeit es ermöglicht, zwischen den Schülern Wissensunterschiede zu erkennen. Es charakterisiert den Test, daß er sich nach Art einer Routine verwenden läßt. Dies gilt in allen drei Testtheorien, doch divergiert die Art der Unterscheidung:
 - Klassische und *probabilistische* Tests lassen sich nur dann anwenden, wenn zwischen den Testpersonen Differenzen auftreten bezüglich des Testmerkmals.
 - Der *kriteriumsorientierte* Test ‚funktioniert‘ auch dann, wenn zwischen den getesteten Personen keine Differenzen auftreten bezüglich des Testmerkmals.Klarer verständlich werden diese Hinweise erst bei Besprechung der drei Testtheorien.
2. Ein Test sieht **standardisierte Situationen** vor: Darin drückt sich zum einen ein theoretisches Ideal, zum anderen ein praktisches Ziel aus - zutreffend für alle drei Ansätze.
 - Das theoretische Ideal betrifft die Objektivier- und Vergleichbarkeit der Daten, es ist ein Meßideal.
 - Das praktische Ziel betrifft die Anwendung: die Absicht, ein Instrument zu entwickeln, das handwerklich, geradezu routinemäßig anwendbar ist.
3. Der Test ‚sammelt‘ eine **Verhaltensstichprobe als Indikator für Personeneigenschaften**. Dies gilt für alle drei Ansätze, aber in unterschiedlichem Sinne:
 - Die *klassische* Testtheorie interpretiert die Personeneigenschaft als relativ stabiles Merkmal.
 - Eine Eigenschaft versteht der *probabilistische* Ansatz ebenfalls als ein stabiles Merkmal, formuliert aber exaktere Annahmen über den Zusammenhang zwischen Testverhalten und diesem Merkmal.
 - Die *kriterienorientierte* Messung faßt das Personenmerkmal offener: Das Merkmal kann einen Verhaltensprozeß betreffen, aber ebenso eine stabile Eigenschaft.
4. Das Testergebnis ermöglicht eine **Zuordnung des Probanden zu einer Gruppe vergleichbarer Personen**. Darin drückt sich eine bestimmte Art der Personbeschreibung aus.

- Gemäß der *klassischen Testtheorie* werden Gruppen von Personen gebildet, deren Testergebnisse vergleichbar sind. Mit den Leistungen solcher Gruppen, der sogenannten Norm- oder Eichstichproben, wird die Testleistung eines einzelnen Probanden oder einer Probandengruppe verglichen.
- Bei der *kriterienorientierten Leistungsmessung* wird eine Testleistung eines Probanden oder einer Gruppe verglichen mit einem Kriterium, das ein bestimmtes Verhalten inhaltlich umschreibt.
- Bei dem *probabilistischen Modell* wird eine Testleistung verglichen mit einer Verteilung von Parametern, die an einer Stichprobe ermittelt und auf Modellverträglichkeit geprüft worden sind.

Für alle drei Ansätze dürfte eine Umschreibung gelten, die besagt:

- ***Test bezeichnet ein diagnostisches Prüfverfahren, das Verhalten in standardisierten Situationen erhebt und Vergleiche mit Gruppen und/oder mit Kriterien ermöglicht.***

Ergänzungshalber zitiert Kasten II-1 eine Definition aus einem klassischen Lehrbuch.

Kasten II-1:

Definition des Tests in einem klassischen Lehrbuch

Quelle: Testaufbau und Testanalyse von Lienert & Raatz (1994, 1).

- „Ein Test ist
- ein wissenschaftliches Routineverfahren
 - zur Untersuchung eines oder mehrerer **empirisch abgrenzbarer Persönlichkeitsmerkmale**
 - mit dem Ziel einer möglichst **quantitativen Aussage**
 - über den **relativen Grad** der individuellen Merkmalsausprägung.“

Einteilung psychodiagnostischer Verfahren

Tests sind unter vielen Perspektiven klassifiziert worden. Eine Einteilung, die einhellig akzeptiert wäre, dürfte es nicht geben.

Um jedoch die Vielfalt der Verfahren zu veranschaulichen, die den Titel ‚Test‘ tragen, sei eine Einteilung von Brickenkamp (1975, 13) übernommen, der generell bei den psychologischen Verfahren drei Hauptklassen unterscheidet - Leistungstests, psychometrische Persönlichkeitstests und Persönlichkeits-Entfaltungsverfahren oder projektive Verfahren:

- **Leistungstests** sind Verfahren, die nach einer Testtheorie konstruiert werden, sie lassen sich charakterisieren durch das Stichwort ‚Performanz‘. Sie verlangen eine Realisierung jenes Verhaltens, das gemessen werden soll. Das Verhalten, das zu realisieren ist, kann unterschiedlichen Bereichen entstammen. Nach diesen verschiedenen Bereichen lassen sich Klassen von

Leistungstests benennen, beispielsweise: Intelligenztests, Konzentrations-tests, Entwicklungstests, Schultests (vgl. Kapitel 9, S. 263).

- **Psychometrische Persönlichkeitstests** werden ebenfalls nach einer Testtheorie konstruiert, sie lassen sich kennzeichnen durch das Stichwort ‚Deskription‘. Sie fordern vom Probanden eine formalisierte Selbstbeschreibung, die sein typisches Verhalten wiedergeben soll. Andere Bezeichnungen lauten: *psychometrische Fragebogen, Persönlichkeitsinventare oder Questionnaire*. Die Beschreibungsdimensionen können unterschiedliche Bereiche betreffen, demnach kann man beispielsweise unterscheiden: Persönlichkeitsstrukturtests, Einstellungs- und Interessentests sowie Klinische Tests (vgl. Kapitel 10, S.263).
- **Persönlichkeits-Entfaltungsverfahren** oder **projektive Verfahren** werden nicht nach einer Testtheorie konzipiert. Sie bezeichnen Instrumente, die den Probanden auffordern, relativ unstrukturiertes Material zu deuten oder zu gestalten. In die Gestaltungen oder Deutungen, so wird angenommen, verlegt der Proband seine Bedürfnisse, Wünsche, Vorstellungen, Fertigkeiten oder Fähigkeiten. Ein mehrstufiger Auswertungsprozeß soll dem Untersucher helfen, diese Wünsche, Bedürfnisse, Vorstellungen, Fähigkeiten zu erschließen (vgl. Kapitel 11, S. 317).

Resümee: Die kurze Charakteristik zeigt, daß nur Leistungs- und Persönlichkeitstests zu der Klasse von Verfahren gehören, auf die eine Testtheorie Anwendung findet. Die Persönlichkeits-Entfaltungsverfahren, die projektiven Verfahren, gehören zu einer eigenständigen Gruppe diagnostischer Vorgehensweisen.

Im Dienste einer eindeutigen **Sprachregelung** sei darum festgelegt: *Wenn im weiteren Fortgang dieses Buches von Tests die Rede ist, sind Leistungs- oder Persönlichkeitstests gemeint.* Für Persönlichkeits-Entfaltungsverfahren oder projektive Verfahren werden wir das Wort Test nicht verwenden.

Eine **Ausnahme** ist angezeigt, wenn das Wort ‚Test‘ zum ‚Eigennamen‘ eines projektiven Verfahrens gehört, wie etwa beim ‚Thematischen Apperzeptions-Test‘ (TAT) von Murray (1943).

Messen als Voraussetzung von Testen

Der Test soll ein Meßinstrument sein, eine Skala, die Maßeinheiten für Verhaltensmerkmale liefert. Was bedeutet das?

„Man kann . . . sagen: Messen bestehe darin, daß wir Objektrelationen, die nicht unsere Erfindung sind, durch Zahlenrelationen abbilden, die unsere Erfindung sind“ (Sixtl, 1967, 3).

Wer demnach eine Messung vornimmt, muß zuerst die Objektrelationen untersuchen, um danach die Zahlenrelationen zu konstruieren. Die Meßtheorie

liefert Kriterien, die zu beurteilen erlauben, wie genau sich die Objektrelationen in den Zahlenrelationen abbilden.

Erinnert sei an die bekannten drei Stichworte, die den Meßvorgang charakterisieren: Repräsentativität, Eindeutigkeit, Bedeutsamkeit.

Repräsentativität betrifft die Qualität der Abbildung: Das empirische Relativ (das Merkmalsgefüge) soll so im numerischen Relativ abgebildet werden, daß die Beziehungen des empirischen Relativs auch in den Beziehungen des numerischen Relativs auftreten.

Eindeutigkeit betrifft die Transformation, die im numerischen Relativ mathematisch möglich ist, ohne die Relationen des numerischen Relativs zu verfälschen.

- Eine Angabe auf **Nominalskalenniveau** bleibt eindeutig, sofern die Zahlen nach der Transformation die gleichen Klassenzuordnungen ermöglichen wie vorher, also die Unterscheidung von Gleichheit und Verschiedenheit. Austauschbar sind die Zahlen, die Gleichheit oder Verschiedenheit ausdrücken. Beispiele sind Haus-, Auto- oder Telefonnummern.
- Eine Skala auf **Ordinalskalenniveau** erlaubt Transformationen, bei denen die Relation von ‚größer/kleiner‘ oder ‚früher/später‘ oder ‚nach/vor‘ gewahrt bleibt. Die Abstände, die zwischen den Objekten im empirischen Relativ bestehen, werden im numerischen Relativ nicht adäquat abgebildet. Beispiele sind Schulnoten.
- Eine Skala auf **Intervallskalenniveau** läßt Transformationen zu, bei denen die Gleichheit der Differenzen in den Einheiten gewahrt bleibt. Die Art der Einheit kann variieren. Beispiele sind Thermometer, Kalenderzeit, man zählt dazu auch Testscores.
- Bei einer Skala auf **Verhältnisskalenniveau** ist eine Transformation zulässig, bei welcher der Abstand zum Nullpunkt exakt angebbar bleibt. Ersetzbar sind die Angabe-Einheiten. Beispiele sind Länge, Gewicht, Zeitangaben, vermutlich auch bestimmte Skalen der Psychophysik.

Bedeutsamkeit betrifft nicht die Frage des semantischen Gehaltes von Skalen, sondern die Frage, welche mathematischen/statistischen Operationen für die einzelnen Skalen erlaubt sind.

Kasten II-2 faßt vier exemplarische Skalentypen zusammen (Bartel, 1971, 12; Bortz, 1989, 31; Stevens, 1963, 25).

**Kasten II-2:
Vier exemplarische Skalenarten**

<i>Skala</i>	<i>Aussagen</i>	<i>Erlaubte Kennwerte Erlaubte Verfahren</i>	<i>Beispiele</i>
Nominal	Gleich/Ungleich	Häufigkeiten/Chiquadrat, Vierfelderkoeffizient	Telefon-, Autonummern
Ordinal	Größer/Kleiner Nach/Vor	Median, Quartile, Percentile/Rangkoeffizient	Windstarken, Schulnoten
Intervall	Gleichheit von Differenzen	Arithm. Mittel, Standardabweichung/ Produktmomentkorrelation	Thermometer, Kalenderdaten
Ratio/Verhältnis	Gleichheiten von Zahlenverhältnissen	Geometrisches Mittel, Varianzkoeffizient	Länge, Gewicht

Gleichgültig, nach welcher Theorie ein Test konstruiert wird, er soll den Anforderungen entsprechen, die in den Sätzen über Messen formuliert sind. Das sei nun für drei testtheoretische Ansätze dargestellt, zuerst für die klassische Testtheorie.

Testtheorien im Dienst von Diagnostik und Intervention

Die drei Testtheorien, die skizziert werden, dienen den beiden Anliegen, Diagnostik und Intervention, in unterschiedlichem Maße:

- Der klassische und der probabilistische Ansatz stehen eher im Dienste der Diagnostik. Aber in begrenztem Maße tragen sie auch zur Intervention bei.
- Der kriteriumsorientierte Ansatz ist eigens zum Zweck der Intervention(smessung) konzipiert worden: vor allem in Pädagogischer und Klinischer Psychologie. Aber auch eine kriteriumsbezogene Messung erbringt zuerst eine diagnostische Aussage.

Die Unterschiede sollten sich bei Darstellung der drei Konzeptionen verdeutlichen.

4. Kapitel

Abriß der klassischen Testtheorie

Die meisten Tests, die heute in psychologischer Diagnostik und Intervention verwandt werden, sind nach den Regeln der sogenannten klassischen Testtheorie konstruiert.

Unter klassischer Testtheorie versteht man ein System syntaktischer Aussagen, an dem sich seit Beginn dieses Jahrhunderts die Konstruktion von Tests orientierte und das 1950 von Gulliksen zusammenfassend formalisiert, 1968 von Lord und Novick erneut überarbeitet und systematisiert worden ist (Michel & Conrad, 1982, 16).

Hier werden nur einige Grundgedanken vorgestellt

Wir skizzieren die Theorie, indem wir die Genese eines *Tests* verfolgen. Unter vier Titeln gibt Kasten 4-1 einen Vorblick auf den schwer eingrenzbaren Lehrstoff (vgl. Tränkle, 1983, 238-240).

Kasten 4-1: Genese eines Tests im Überblick

Entwurf:

- Sichtung theoretischer Ansätze und empirischer Befunde, die zum Thema vorliegen,
- Konzeptualisierung und Ausarbeitung der Fragestellung,
- Festlegung der Merkmale, die erfaßt (oder miterfaßt) werden sollen,
- Befragung von Experten.

Erprobung:

- Vorlage der Testvorform bei einer Stichprobe, die der Zielgruppe möglich ähnlich ist:
 - ⇒ zur Ermittlung mißglückter Itemformulierungen,
 - ⇒ zur Analyse der Testaufgaben (Itemanalyse),
- Auswertung.

Revision:

- Selektion, Elimination oder Überarbeitung der Items aufgrund der Erprobungsergebnisse,
- gegebenenfalls Wiederholung der Erprobung,
- Erstellung eines revidierten Tests.

Endfassung:

Ziehung einer angemessenen Stichprobe der Zielgruppe,

- Vorlage und Auswertung des revidierten Tests:
 - ⇒ Bestimmung von Standardisierung (Objektivität), Meßgenauigkeit (Reliabilität) und Gültigkeit (Validität),
 - ⇒ Festlegung genereller und, sofern möglich, spezieller Normen,
- Interpretation der Ergebnisse unter Berücksichtigung der methodischen Beschränkungen,
- Vergleich mit Ergebnissen, die auf ähnlichen oder die auf andersartigen Methoden beruhen.

Die Darstellung des Kapitels gliedert sich in sechs größere Abschnitte:

- Fragestellung, Testmerkmal, Test-Item (4.1),
- Itemanalyse (4.2),
- Ermittlung der Testgutekriterien (4.3),
- Normierung oder Eichung (4.4),
- Beitrag zu Diagnostik und Intervention (4.5),
- Kritik der klassischen Testtheorie (4.6).

Das Kapitel schließt mit einer Zusammenfassung (4.7) und der Vorgabe einiger Kontrollfragen (4.8).

4.1 Fragestellung, Testmerkmal, Test-Item

Wir besprechen fünf Teilprobleme:

- Konzeptualisierung von Fragestellung und Testmerkmal (4.1.1),
- Zuordnung von Testmerkmal und Test-Item (4.1.2),
- Wahl einer Konstruktionsstrategie (4.1.3),
- Bestimmung der Testart (4.1.4),
- Itemgenerierung und Itemgestaltung (4.1.5).

4.1.1 Konzeptualisierung von Fragestellung und Testmerkmal

Die klassische Testtheorie präsentiert ein formales System. Vor Anwendung ihrer Regeln stellen sich jedoch inhaltliche Fragen, auf welche sie keine Antwort vorsieht. Eine solche Frage betrifft den Grund, warum ein ‚neues‘ Instrument entwickelt werden soll, betrifft also die **Fragestellung**, von der eine Testkonstruktion ausgeht. Die Anlässe können verschieden sein:

- *Ein Untersucher mag in seiner (praktischen oder wissenschaftlichen) Arbeit auf eine Aufgabe stoßen, für die er von fertigen Instrumenten keine Lösung erwartet, für deren Bewältigung er darum ein neues Verfahren anfertigen will, etwa: Wie läßt sich Kreativität bei Sonderschülern messen? Wie läßt sich ‚Lebenszufriedenheit‘ von Menschen in Altersheimen erfassen?*

- *Eine Institution mag ein ‚neues Problem‘ entdecken oder ein altes Problem ‚neu sehen‘ und die Entwicklung eines Instruments in Auftrag geben, von dem sie eine Lösung erhofft: Im Auftrag der Kultusminister der Länder der Bundesrepublik Deutschland wurde der ‚Test für medizinische Studiengänge (TMS)‘ eingeführt. Zusätzlich zum Abitur ist er ein Kriterium bei der Vergabe medizinischer Studienplätze (Trost und Mitarb., 1995; Fay, 1982; Fisseni, Olbrich, Halsig, Mailahn & Ittner 1993).*

Aus der Fragestellung ergeben sich konkrete Regeln und Schritte für die Konstruktion eines ‚neuen Instrumentes‘. Sie sind jeweils spezieller Natur. - Eingeschlossen ist aber eine Frage, die allgemeiner Natur ist: Auf welche Weise ist das Merkmal zu bestimmen, das der ‚neue Test‘ erfassen soll?

Zunächst stellt sich die Aufgabe einer **Abgrenzung** - einer **theoriegeleiteten Definition**. Jede Messung von Merkmalen beruht auf theoretischen Annahmen über Verhalten, über Messen und Meßinstrumente. Bei der Testkonstruktion sollten die Annahmen explizit genannt werden. *Beispiel*: Thurstone (1938) hat aus einer faktorenanalytisch gewonnenen Strukturierung kognitiver Dimensionen Verhaltensbereiche ‚ausgegrenzt‘, die ein neuer Test erfassen sollte (Primary mental abilities).

*Das **Merkmal**, das der Test ‚abbilden‘ soll, sei umschrieben*

- als eine Zusammenfassung (als ein Kürzel) mehrerer empirisch beobachtbarer Verhaltensweisen, die
- eine gewisse Konstanz über Zeiträume (die nicht definiert sind) und
- eine gewisse Konsistenz über Situationen (die relevant sind für das Verhalten) aufweisen sollen.

Anschaulicher schreibt Klauer (1987, 13-14):

*„Unter **Persönlichkeitsmerkmal** verstehen wir die Eigenschaft eines Menschen, in einer gegebenen Klasse von Situationen eine bestimmte Klasse von Verhaltensweisen zu äußern.“*

Damit ist auch gesagt: Das Merkmal, das ein Test erfassen soll, ist nicht beobachtbar, es ist nur erschließbar. Es ist **kein Beobachtungs-**, es ist ein **Dispositionsprädikat**. Die ‚Verhaltensweisen‘ dagegen, welche den Schluß ermöglichen, müssen beobachtbar sein: etwa die Aufgabenlösungen in einem Leistungstest und die Antworten auf Fragen in einem Persönlichkeitsinventar.

Der folgende Abschnitt soll die Herleitung konkreter Konstruktionsschritte aus theoretischen Annahmen an einem Beispiel veranschaulichen.

Demonstration:

Eine theoriegeleitete Testkonstruktion.

Entwicklung der ‚Personality Research Form‘ (PRF)

Der Zusammenhang zwischen theoretischem Ansatz und Konstruktion eines Fragebogens sei veranschaulicht an der Herleitung des Fragebogens ‚Personal-

ity Research Form‘ aus Murrays Personologie (Angleitner, Stumpf & Wieck, 1976; Jackson, D.N. 1974; Stumpf et al., 1985):

Ansatz: Jackson, der Autor der PRF, ging von Murrays Theorie über Bedürfnis und Druck (need and press) als (motivationale) Ursprünge des Verhaltens aus.

Merkmalsdefinition: Er übernahm die Konstrukte der ‚Bedürfnisse‘ (nicht die der ‚Drucke‘), definierte sie aber alle neu.

Beispiel: *Autonomie in hoher Ausprägung umschreibt ein Verhalten, das darauf abzielt, sich freizuhalten von Beschränkungen jeder Art, Aktivitäten zu vermeiden, die eine höhere Autorität vorschreibt, nach Freiheit und Unabhängigkeit zu streben, den eigenen Eingebungen zu folgen und Konventionen zu umgehen (vgl. Angleitner Stumpf & Wieck, 1976, 18; Stumpf et al., 1985, 44).*

Formulierung von Testfragen: An den ‚Definitionen orientierte sich die Formulierung der Testfragen, der sogenannten ‚Items‘.

Beispiele für die Skala ‚Autonomie‘ (Angleitner et al., 1976, 143, 146, 147 Fragebogenform AA):

- *Ich würde gern frei durch die Länder ziehen.* (Item 27)
- *Abenteuer die ich allein durchstehen muß,
beängstigen mich ein bißchen.* (Item 49)
- *Ich will vor allem unabhängig und frei sein.* (Item 115)
- *Ich versuche meistens, meine Sorgen mit jemandem zu teilen,
der mir helfen kann.* (Item 137)

Zwei Experten überprüfen, ob die Testfragen folgenden Kriterien entsprachen (Stumpf et al., 1985, 9):

- „ 1. Konformität zum zugehörigen Konstrukt,
2. angemessene Repräsentierung positiver wie negativer Ausprägungen eines jeden Merkmals,
3. Klarheit und Unzweideutigkeit,
4. Freiheit von extremer sozialer Erwünschtheit (bzw. Unerwünschtheit),
5. voraussichtliche Diskriminationsfähigkeit und hinreichende Popularität in den in Frage kommenden Bezugsgruppen,
6. Repräsentativität der einzelnen Itemmengen in bezug auf das jeweilige Merkmal.“

Personenbeschreibung: Jackson beschrieb (fiktive) Personen mit hoher und Personen mit niedriger Ausprägung der drei Merkmale ‚Autonomie‘, ‚Impulsivität‘ und ‚Dominanz‘. Zu jedem wurden sechs Items vorgegeben.

59 Beurteiler sollten angeben, „für wie groß sie die Wahrscheinlichkeit hielten, daß die fiktiven Personen dem Item beipflichten wurden“ (Stumpf et al., 1985, 9).

Bei dem folgenden Beispiel handelt es sich um die Beschreibung einer Person, die eine geringe Ausprägung von Autonomie besitzt (Stumpf et al., 1985, 9):

„Alex Reed arbeitet im Werbebüro einer großen Versicherungsgesellschaft. Obwohl er für die Werbeauslagen seiner Firma verantwortlich ist, sucht er stets den Rat seiner Mitarbeiter und Vorgesetzten, bevor er eine Entscheidung trifft. Normalerweise hält er eine Versammlung ab, in der sich jeder über die geplanten Werbemaßnahmen äußern kann, und man kommt dann zu einer Gruppenentscheidung, die er seinem Vorgesetzten zur Genehmigung vorlegt. Alex macht es besondere Freude, sein Büro als Team zu organisieren, in dem die einzelnen in allen Arbeitsabschnitten zusammenarbeiten.“

Testanalyse: Der Fragebogen wurde Probanden vorgelegt, die Antworten - zur Ermittlung der Gütekriterien - den üblichen Analysen unterworfen.

4.1.2 Zuordnung von Testmerkmal und Test-Item

Ein Test, der nach der klassischen Testtheorie konstruiert wird, erfaßt ein Merkmal durch Fragmentierung: Kleine Einheiten, die sogenannten Test-Items, sollen das Testmerkmal ‚inhaltlich‘ repräsentieren. Für die Beantwortung von Items werden Punkte gegeben: die Item-Scores. Die Summe dieser Punkte ist der Test-Score: quantitativer Repräsentant der Ausprägung des Testmerkmals. Kasten 4.1-1 bietet eine Übersicht.

Kasten 4.1-1:
Test-Item, Item-Score, Test-Score

Test-Item:	<i>Kleinste Einheit in einem Test, Einzelaufgabe oder Einzelfrage</i>
Item-Score:	<i>Punktwert für die Beantwortung eines Items, etwa 1 für ‚Richtig‘, 0 für ‚Falsch‘</i>
Test-Score oder Summen-Score:	<i>Summe der Item-Scores</i>

Mit dem Problem der Merkmals-Erfassung durch Items verschränkt sich eine andere Frage: die der Konstruktionsstrategien.

4.1.3 Wahl einer Konstruktionsstrategie

Welche Items sollen zu einem Test zusammengefaßt werden? Um diese Frage zu beantworten, empfehlen sich drei Entscheidungsstrategien (Goldberg, 1971; Hase & Goldberg, 1967):

Die **rationale Strategie** besteht darin, von einem vorgegebenen theoretischen Konzept her Items zu formulieren und sie dann weiteren Prüfungen zu unterziehen.

Beispiel: Horn (1983) konstruierte das ‚Leistungsprüfsystem‘ (LPS), indem er einer Entwicklung von Items und einer Gruppierung der Subtests die Intelligenztheorie von Thurstone zugrunde legte.

Die *extemale Strategie* besteht darin, Items zu einem Test zusammenzustellen, die zwischen einer Kriteriumsgruppe und einer Kontrollgruppe unterscheiden. Die Aussagen des Tests ergeben sich aus dem Merkmal, das die Kriteriumsgruppe charakterisiert (und das die Kontrollgruppe nicht besitzt). Der Inhalt der Items selber ist demnach irrelevant. Auch unsinnige oder subtile Items bleiben in einer Skala, wenn sie Kriteriums- und Kontrollgruppe trennen.

Beispiel: Konstruktion des *Minnesota Multiphasic Personality Inventory‘* (MMPI: Dahlstrom, Welsh & Dahlstrom, 1972). Seine Autoren, Hathaway und McKinley, legten Gruppen von psychiatrisch Kranken (Kriteriumsgruppen) und Gruppen von Normalen (Kontrollgruppen) Items vor. Zu einer Skala wurden solche Items zusammengefaßt, deren Mittelwerte zwischen beiden Gruppen signifikante Unterschiede aufwiesen.

Die *internale Strategie* besteht darin, daß ein Itemsatz (theoretisch oder atheoretisch zusammengestellt) einer Stichprobe vorgelegt wird, dann solche Items zu einer Skala zusammengezogen werden, die eine statistische Prozedur als zusammengehörig erweist. Den Inhalt einer Skala bestimmt der Inhalt der Items, welche zur Skala zusammentreten.

Beispiel: Konstruktion des ‚Freiburger Persönlichkeitsinventars‘ (FPI: Fahrenberg, Selg & Hampel, 1978; FPI und FPI-R: Fahrenberg, Hampel & Selg, 1989). Ein Itempool wurde zwei Stichproben vorgelegt. Die Test-Scores wurden einer Faktorenanalyse unterzogen. Items, die auf demselben Faktor hoch luden, wurden zu einer Skala gruppiert. Benannt wurden die Skalen nach dem Inhalt ihrer Items.

Vergleich: In empirischer Sicht hat sich ergeben: Alle drei Strategien führen zu ähnlichen Ergebnissen, zu Skalen, die brauchbare psychometrische Qualitäten aufweisen (Angleitner, 1976, 39; Hase & Goldberg, 1967; Jackson, D. N., 1975). Aus theoretischer Sicht bleibt es jedoch unbefriedigend, Merkmale allein durch statistische Verfahren oder allein über Kriteriumsgruppen zu bestimmen. Am Beginn einer Testkonstruktion sollte darum eine theoretische Abgrenzung stehen, ihr entspricht als Konstruktionsstrategie die ‚rationale Vorgehensweise‘. Die ‚extemale‘ und die ‚internale‘ Strategie sollten danach der Prüfung dienen, ob empirische Befunde die theoretischen Abgrenzungen stützen.

4.1.4 Bestimmung der Testart

In Wechselbeziehung zu den Konstruktionsstrategien und zur Konzeptualisierung des Testmerkmals steht eine dritte Frage: die Festlegung der Testart.

Zunächst geht es um pragmatische Fragen:

- Einzel- oder Gruppentest?
- Material: nur Papier und Bleistift oder auch apparative Ausstattungen?
- Testlänge?
- Für welche Population ist der Test vorgesehen, wie lassen sich die entsprechenden Stichproben gewinnen?
- Struktur: Ein einziger Gesamttest oder Gliederung in Untertests?

Schon solche pragmatischen Fragen sollten bei den Vorüberlegungen bedacht und entschieden werden. Im Hintergrund bleiben weitere Probleme zu lösen:

- Bei der Frage nach **Einzel- oder Gruppentests** ist zu klären, ob Messung in einer (partnerschaftlichen?) Zweisituation zu gleichen Ergebnissen führt wie Messung in einer (distanzierteren?) Gruppensituation.
- Bei der Frage nach **Papier- oder Apparate-Tests** ist zu klären, ob sich Merkmale in beiden Medien auf gleiche Weise ‚abbilden‘. Läßt sich etwa ‚technische Begabung‘ gleich gut messen mit Fragen, die auf dem Papier zu beantworten sind, wie mit entsprechenden Apparaten?
- Was die **Testlänge** betrifft, so geht in diese Festlegung auch ein, ob man differenzierte oder nur globale Meßergebnisse erwartet. Der ‚Reduzierte Wechsler-Intelligenztest‘ (WIP: Dahl, 1972) beschafft nicht so viele Informationen wie der vollständige ‚Hamburg-Wechsler-Intelligenztest‘ (HA-WIE-R: Tewes, 1991), aber möglicherweise liefert er ausreichende Informationen für bestimmte Fragestellungen.
- Was die **Population** angeht, für die ein Test vorgesehen wird, so gilt: Je spezieller die Selektion, je homogener die angezielte Gruppe, um so schwieriger die Ermittlung ausreichender Gütekriterien. Je globaler dagegen die Population, je heterogener ihre Zusammensetzung (bezogen auf das Testmerkmal), um so leichter ist es, günstige Gütekriterien zu gewinnen.
- Wenn schließlich zu entscheiden ist, ob eine **einzelne Gesamt- oder eine differenzierte Strukturaussage** angestrebt wird, so liegt hier der Bezug zur Persönlichkeitspsychologie besonders nahe; denn ‚Strukturen‘ sollten nicht bloß pragmatisch konzipiert, sondern theoretisch begründbar sein.

Festzulegen, welche Testart konstruiert werden soll, verschränkt sich darum mit der (vorher beschriebenen) Aufgabe, zu entscheiden, ob das Testmerkmal durch eine theoretische Strategie abgegrenzt (definiert) und in statistisch-empirischen Strategien getestet werden soll, aber ebenso mit der gesamten Problematik einer Merkmalsdefinition.

Demgegenüber nimmt sich der nächste Schritt weniger theoretisch aus - allerdings nur auf den ersten Blick.

4.1.5 Itemgenerierung und Itemgestaltung

Die Generierung von Items, aus denen sich ein Test aufbaut, erweist sich als eine sehr komplexe Arbeit (Aiken, 1982; Heidenreich, 1989; Klauer, 1987; Lienert & Raatz, 1994; Roidt & Haladyna, 1982).

Am schwierigsten ist das Problem zu lösen, wie sich die Items streng theoretisch ableiten und regelhaft ‚vervielfältigen‘ lassen.

Hier sei nur eine Frage herausgegriffen, die der semantischen Gestaltung. Rütter (1978) gibt einen reichhaltigen Überblick über die Vielfalt der Gestaltungsmöglichkeiten.

Wir besprechen drei Aspekte:

- 1. Aufgaben nach Antwortart (gebunden oder frei),
- 2. Aufgaben nach Inhaltsumfang (einfach oder komplex),
- 3. Aufgaben nach Darstellungsmedium (verbal oder nichtverbal).

(1) Aufgaben nach Art der Antwort:
Gebundene oder freie Items

a) Gebundene Items

Als gebunden bezeichnet man Aufgaben, die dem Probanden ein Problem stellen, ihm zugleich aber verschiedene ‚Lösungen‘ anbieten. Ein Schema soll einen Überblick geben (Heidenreich, 1989, 401).

Gebundene Items			
Auswahlaufgaben		Ordnungsaufgaben	
Richtig-falsch-Aufgaben	Mehrfachwahl-aufgaben	Zuordnungs-aufgaben	Umordnungs-aufgaben

Vorteile dieses Aufgabentyps sind die eindeutige Vergleichbarkeit der Antworten und die Möglichkeit routinemäßiger Auswertung. Ein Nachteil liegt darin, daß Items dieser Art eher reaktives als kreatives Verhalten erfassen.

Auswahlaufgaben:

Richtig-Falsch-Aufgaben: Das Problem wird dargestellt, es werden zwei Antwortalternativen geboten, nur eine ist richtig.

Beispiel: *Eine Quadratzahl wird gebildet, indem eine Zahl mit sich selbst multipliziert wird. – Richtig []. Falsch [].*

Mehrfachauswahl (multiple choice): Das Problem wird dargestellt, zugleich werden mehrere Antwortalternativen angeboten, von denen nur eine richtig ist.

Beispiel: IST 70¹, Satz-Ergänzen: *Quecksilber ist ein/eine...? a) Metall, b) Mineral, c) Lösung, d) Gemisch, e) Legierung. (Richtige Antwort: a)*

HINWEIS: Die Mehrfachauswahl ist der dominierende Aufgabentypus heutiger Tests.

Ordnungsaufgaben:

Zuordnungsaufgaben: Problem und Lösung einer Aufgabe werden vorgegeben, der Proband soll beide einander zuordnen.

Beispiel:	<i>Zu welcher Baumform gehören?</i>	<i>Lösung ankreuzen</i>
(a) Linde	(1) Pyramidenbaum	(a) 1 2 3 4
(b) Trauerweide	(2) Schirmbaum	(b) 1 2 3 4
(c) Akazie	(3) Kugelbaum	(c) 1 2 3 4
(d) Fichte	(4) Hängebaum	(d) 1 2 3 4

(Lösungen: a = 3; b = 4; c = 2; d = 1)

Umordnungsaufgaben: Elemente einer Aufgabe werden ungeordnet vorgegeben, der Proband soll sie ordnen.

Beispiel: *Sucht Dame Hausfrau als junge Anstellung.*
(Lösung: Junge Dame sucht Anstellung als Hausfrau.)

b) Freie Items

Als frei bezeichnet man Aufgaben, die das Problem vorgeben, aber keine ‚Lösung‘ anbieten. Der Proband muß die Lösung selber formulieren. Wieder soll ein Schema einen Überblick geben (Heidenreich, 1989, 401).

Freie Items	
<i>Ergänzungsaufgaben</i>	<i>Kurzantwort, Kurzaufsatz</i>

Der Vorteil dieses Aufgabentyps liegt darin, daß er die Erfassung einer großen Verhaltensbreite erlaubt. Der Nachteil besteht darin, daß es schwer ist, unterschiedliche Antworten gleich zu gewichten.

Ergänzungsaufgaben: Das Problem wird vorgegeben, der Proband muß die Frage *kurz* ergänzen. Vielfältige Formen sind möglich.

Beispiele:

- *Wo liegt Ägypten?*
- *Apfel verhält sich zu Obst wie Weizen zu ...?*

Kurzantwort, Kurzaufsatz: Ein Problem wird vorgegeben, der Proband erhält die Möglichkeit, längere Antworten zu geben. Auch hier sind vielerlei Formen möglich.

¹ IST 70: „Intelligenz-Strukturtest 70“ von Amthauer (1973).

Beispiele:

- Warum wächst dieselbe Baumart in Finnland langsamer als in Italien?
- Gestern traf ich auf dem Marktplatz einen Schulkollegen, den ich zwanzig Jahre nicht gesehen hatte. Da...

**(2) Aufgaben nach dem Inhaltsumfang:
Einfache oder komplexe Aufgaben**

Nach ihrem Inhalt behandeln Aufgaben einfache oder komplexe Probleme:

- **Einfache Items** setzen wenig Vorwissen voraus, erfordern eine einfache Stellungnahme.
- **Komplexe Items** setzen mehr Vorwissen voraus, die Lösungswege sind nicht so überschaubar.

Die Einteilung in ‚einfache und komplexe Items‘ ist stichprobenbezogen: So könnte die Ergänzungsaufgabe, die eben zitiert wurde (*Apfel verhält sich zu Obst wie Weizen zu...?*), für Sonderschüler zu komplex, für Gymnasiasten zu einfach sein.

Fingerspitzengefühl ist vonnöten, um Items von mittlerer Komplexität zu formulieren: Items, die nicht so leicht sind, daß alle sie lösen, aber auch nicht so schwer, daß keiner eine Lösung findet.

**(3) Aufgaben nach dem Darstellungsmedium:
Verbal oder nichtverbal**

Die Einteilung nach dem Darstellungsmedium klassifiziert die Items (vor allem) danach, ob sie in Worte gefaßt oder nonverbal in Bildern und Verhaltensprozessen dargestellt werden.

Beispiele: Der ‚*Figure Reasoning Test*‘ (FRT) gibt Aufgaben in symbolischer Kodierung (Daniels, 1971) der HAWIE gibt sowohl verbale wie nonverbale Items vor

Sprache ist wichtigstes Mittel der Verständigung, doch setzt sie auch Barrieren. Nichtverbale Items sollen die Barrieren senken, können sie aber nicht beseitigen. Denn solche Items bedürfen ihrerseits der sprachlichen Interpretation.

4.2 Itemanalyse

Die Itemanalyse soll prüfen, ob die Test-Items der Test-Absicht entsprechen. Die ‚Entsprechung‘ wird vor allem durch drei Gütekriterien geprüft:

- Schwierigkeitsindex (4.2.1),
- Trennschärfe (4.2.2),
- Homogenität (4.2.3).

Nach Abschluß der Itemanalyse ist zu entscheiden
über die Selektion geeigneter Items (4.2.4).

4.2.1 Schwierigkeitsindex

Der Schwierigkeitsindex oder ‚Index der kategorialen Häufigkeiten‘ gibt an, wie groß der Anteil von Probanden ist, die ein Item ‚richtig‘ beantwortet haben (Michel & Conrad, 1982, 20).

Die Frage, die der Schwierigkeitsindex beantworten soll, ergibt sich aus dem differentialpsychologischen Ansatz der klassischen Testtheorie: Ein Test soll Probanden mit hoher Merkmalsausprägung **trennen** von Probanden mit geringer Merkmalsausprägung. Zu einer solchen Unterscheidung tragen zwei Klassen von Items nichts bei: erstens solche Items, die von **allen** Probanden, zweitens solche, die von **keinem** Probanden ‚gelöst‘ werden. Der Schwierigkeitsindex soll Items identifizieren, die ‚brauchbarer‘ sind als diese zwei Klassen.

Wir besprechen folgende Fragen:

- Schwierigkeitsindex bei zweistufigen Antworten (4.2.1.1),
- Schwierigkeitsindex bei mehrstufigen Antworten (4.2.1.2),
- Erwünschte Schwierigkeitsindizes (4.2.1.3),
- Schwierigkeitsindex und andere Itemkennwerte (4.2.1.4).

4.2.1.1 Schwierigkeitsindex bei zweistufigen Antworten

Bei Items, die eine zweistufige Antwort erfordern (Ja/Nein; Richtig/ Falsch), berechnet sich der Schwierigkeitsindex (p) als Quotient aus ‚Zahl der Richtiglöser‘ (N_R) und ‚Zahl der Probanden‘ (N) (Lienert & Raatz, 1994, 74-78):

$$p = \frac{N_R}{N}$$

Ein Item, das von vielen gelöst wird, erhält ein hohes p , es ist leicht. Ein Item, das von wenigen gelöst wird, erhält ein niedriges p , es ist schwer. *Die semantische Bedeutung (leicht, schwer) und die quantitative Repräsentation (hoher, niedriger Wert) sind einander also gegensinnig zugeordnet.*

Haben nicht alle Probanden alle Items beantwortet, dann gibt N eine falsche Bezugsgröße an: p fällt zu niedrig aus ($p \Rightarrow$ ‚schwerer‘). Um diese Verzerrung aufzufangen, ist es möglich, in die Berechnung nur jene Probanden einzubeziehen, die ein Item beantwortet haben (N_B):

$$p = \frac{N_R}{N_B}$$

Kasten 4.2-1 gibt (fiktive) Werte für ein Berechnungsbeispiel. Bei Item 1 steht im Nenner die Probandenzahl (N), bei Item 4 die Zahl der Probanden, die Item 4 bearbeitet haben (Nu).

$$p_1 = N_{R1}/N = 4/7 = .57$$
$$p_4 = N_{R4}/N_{B4} = 4/5 = .80$$

Kasten 4.2-1:
Berechnung von Schwierigkeitsindizes bei zweistufigen Antworten

1, 0 : Item gelöst, nicht gelöst : Item nicht beantwortet N_R : Zahl der Richtiglöser	N_B : Zahl der Pbn, die ein Item beantwortet haben N : Zahl der Pbn, die beteiligt waren									
	Items									
Pbn	1	2	3	4	5	6	7	8	9	10
1	1	1		1			1		n	
2	0	0		1			1		n	
3	1	0		1			1		0	
4	0	0		1			1		0	
5	1	0		0			1		0	
6	0	0		n			1		0	
1	1	1		n			1		0	
N_R	4	2		4			7		0	
N_B	7	7		5			1		5	
N	7	7		7			7		7	
p	.57	.28		.80			1.00		0.00	

Zufallskorrektur

Bei Leistungstests, deren Lösungen zweistufig kodiert werden, kann **Erraten** erheblichen Einfluß auf die Zahl der Richtiglösungen ausüben: Fünfzig Prozent ‚richtiger‘ Antworten **können** auf Raten zurückgehen. Für diesen Fall hat Guilford (1936) eine Formel vorgeschlagen, die den Zufallseinfluß reduzieren soll.

$$p_c = \frac{N_R - \frac{N_F}{m - 1}}{N}$$

- Neu sind die Terme:
- P_c : Korrigierter Schwierigkeitsindex,
 - N_F : Falschlöser,
 - m : Zahl der Alternativen in einer Aufgabe.

(Bei Ja-Nein-Aufgaben gibt es zwei Alternativen, bei Mehrfachauswahl kann es beliebig viele Alternativen geben.)

Aus Kasten 4.2-1 sei der Index von Item 1 in dieser Weise ‚korrigiert‘, die Zahl der Falschlöser (NF) beträgt 3, die Zahl der Alternativen (m) betrage 4:

$$p_{c,1} = \frac{4 - \frac{3}{4-1}}{7} = 0.43$$

Der korrigierte Schwierigkeitsindex für Item 1 (0.43 gegenüber 0.57) besagt, daß Item 1 schwieriger ist, als es zunächst erscheint, jedenfalls dann, **wenn** man eine Zufallskorrektur gelten läßt.

4.2.1.2 Schwierigkeitsindex bei mehrstufigen Antworten

Der Schwierigkeitsindex p ist definiert für zweistufige Antworten. Sehen Item-Antworten mehrere Abstufungen vor, etwa Werte von 0 bis 10, dann ist p zur Angabe der ‚Schwierigkeit‘ nicht definiert. Um dennoch Angaben zu gewinnen, kann man unterschiedlich vorgehen:

- Man kann die Item-Scores an einem Kriterium dichotomisieren und gibt den Werten unterhalb eine 0, den Werten oberhalb eine 1. Auf diese Weise kann man das ‚zweistufige‘ p berechnen. Allerdings verzichtet man auf Differenzierungen, die in den Abstufungen enthalten sind.
- Man berechnet Mittelwert und Streuung je Item und verwendet den Mittelwert als Äquivalent zu p. (p ist seinerseits ein Mittelwert.) Allerdings ist ein Mittelwert ohne gleichzeitige Beachtung des Streuungsmaßes wenig aussagekräftig.
- *Man berechnet einen eigenen Wert für p: Er sei ‚mehrstufiges p‘ genannt, sein Kürzel sei p_m .*

Um den dritten Vorschlag, die **Berechnung eines ‚mehrstufigen p‘ (p_m) zu demonstrieren**, sei von einem Beispiel ausgegangen.

Beispiel: Beim HAWIE, Untertest ‚Figuren-Legen‘ (FL), kann man zur Aufgabe 3 (eine Hand zusammenlegen) 0 bis 10 Punkte erreichen. 10 Probanden seien mit dem FL getestet worden. Wieviel Punkte können alle zusammen maximal erreichen? 10 Probanden können maximal 10 x 10 Punkte = 100 Punkte erreichen. Tatsächlich seien weniger Punkte erreicht worden:

- 3 Probanden erreichen 4 Punkte, zusammen 12;
- 4 Probanden erreichen 6 Punkte, zusammen 24;
- 3 Probanden erreichen 9 Punkte, zusammen 27.

Zusammen erreichen die 10 Probanden 63 Punkte. Es stehen zur Berechnung demnach zur Verfügung

- **erreichte** Wertpunkte (hier: 63) und
- **erreichbare** Wertpunkte (hier: 100).

In Analogie zum zweistufigen Schwierigkeitsindex läßt sich ein Index für mehrstufige Antworten (p_m) bestimmen, der die beiden genannten Werte in Beziehung setzt:

$$p_m = \frac{\text{Erreichte Wertpunkte}}{\text{Erreichbare Wertpunkte}}$$

$$p_m = \frac{63}{100} = 0.63$$

Von drei Autoren liegen Vorschläge vor, p_m in der beschriebenen Weise zu berechnen: von Dahl (1971) sowie von Wagner und Baumgärtel (1978). Zwar unterscheiden sich die Vorschläge im einzelnen, doch laufen sie auf gleiche Ergebnisse hinaus:

$$p_m = \frac{\Sigma X}{\Sigma X_{\max}} = \frac{\Sigma X}{N \cdot X_{\max}}$$

Es bedeuten:

- p_m : Schwierigkeitsindex für mehrstufige Items,
- X : Item-Score,
- X_{\max} : Maximaler Item-Score (X läuft von 0 bis X_{\max}),
- ΣX_{\max} : Summe von X_{\max} über alle Probanden,
als Äquivalent gilt: $N \cdot X_{\max}$.

Kasten 4.2-2 gibt (fiktive) Ergebnisvektoren für zwei Items.

Kasten 4.2-2:
Schwierigkeitsindex bei mehrstufigen Items

X : Item-Score (X reicht von 0 bis 5)		$f_1 \ f_2$: Lösungs-Häufigkeit bei Item 1 und Item 2						p_m : Schwierigkeitsindex für mehrstufige Items	
ΣX : Test-Score									
X	\Rightarrow	0	1	2	3	4	5	sx	Pm
Item 1	f_1	1	3	2	5	2	1	35	0.50
2	f_2	7	0	0	0	0	7	35	0.50

Mit den Werten aus Kasten 4.2-2 sei p_m für Item 1 und Item 2 berechnet:

$$\begin{aligned} N_1 &= N_2 = 14 \\ X_{\max 1} &= X_{\max 2} = 5 \\ \Sigma X_1 &= 1 \times 0 + 3 \times 1 + 2 \times 2 + 5 \times 3 + 2 \times 4 + 1 \times 5 = 35 \\ \Sigma X_2 &= 7 \times 0 + 0 \times 1 + 0 \times 2 + 0 \times 3 + 0 \times 4 + 7 \times 5 = 35 \\ \Sigma X_{\max 1} &= 14 \times 5 = 70 = \Sigma X_{\max 2} \end{aligned}$$

Die Schwierigkeitsindizes für Item 1 und Item 2 fallen gleich aus:

$$p_{m,1} = 35/70 = 0.50 = p_{m,2}$$

Einwand: Nun läßt sich folgender Einwand erheben: Die zwei Items nehmen denselben Wert für p_m an, obwohl ihre Punkteverteilungen erheblich divergieren. Zwei Items mit unterschiedlichen Verteilungen haben unterschiedliche Va-

rianzen. Nun gilt aber: „Je größer die Itemvarianz, um so unterschiedlicher reagieren die Personen auf das Item und um so besser kann man mit ihm differenzieren“ (Kranz, 1981, 54). Haben demnach zwei Items den gleichen Wert für p_m , aber unterschiedliche Varianzen, dann differenziert das Item mit der größeren Varianz eindeutiger als das Item mit der geringeren Varianz.

Diesen Unterschied kann p_m nicht wiedergeben. Der Grund liegt darin, daß in die Berechnung nur der Mittelwert eingeht, nicht die Varianz. (Bei dem zweistufigen p ist mit dem Mittelwert, p , die Streuung mitbestimmt.)

Die **Varianzen** (s^2) in dem Beispiel divergieren aber erheblich:

$$s^2_1 = 1.82 \neq 6.25 = s^2_2$$

Ergänzungsvorschlag: Zu wünschen wäre eine Berechnung, in welche die Unterschiede der Varianzen eingehen. Eine Formel, welche die gewünschte Unterscheidung ermöglicht, könnte bei den Vorschlägen von Dahl, Wagner und Baumgärtel ansetzen, aber - in Analogie zur Bildung der Varianz - statt der einfachen Punktwerte ihre Quadrate einbeziehen. Sie könnte lauten:

$$p_m = \frac{\sum X^2}{\sum X^2_{\max}}$$

Der Ergänzungsvorschlag sei veranschaulicht an den Daten von Kasten 4.2-2:

$$\begin{aligned} \sum X^2_1 &= 1 (0^2) + 3 (1^2) + 2 (2^2) + 5 (3^2) + 2 (4^2) + 1 (5^2) \\ &= 113 \\ \sum X^2_2 &= 175 \\ \sum X^2_{\max 1} &= 14 (5^2) = 350 = \sum^2_{\max 2} \\ p_{m,1} &= 113/350 = 0.32 \\ p_{m,2} &= 175/350 = 0.50 \end{aligned}$$

Die Schwierigkeitsindizes für Item 1 und Item 2 divergieren:

$$p_{m,1} = 0.32 \neq 0.50 = p_{m,2}$$

Die Ergänzungsformel wurde zu folgenden Resultaten führen:

- Je mehr Probanden den maximalen Item-Score erreichen (X_{\max}), desto mehr nähert sich p_m dem Wert 1 (leichte Items).
- Je mehr Probanden den minimalen Item-Score erhalten (0), desto mehr nähert sich p_m dem Wert 0 (schwere Items).
- Wenn die Varianz von Item i ungleich ist der Varianz von Item j, dann folgt: $p_{m,i}$ ist ungleich $p_{m,j}$.
- Bei maximaler Varianz eines Items erreicht p_m den Wert von 0.50 - analog dem Schwierigkeitsindex p für dichotome Daten.
- Die Formel enthält den zweistufigen Schwierigkeitsindex als Sonderfall, nämlich als den Fall, in dem gilt $X_{\max} = 1$. Insofern sind beide Indizes ineinander konvertierbar.

4.2.1.3 Erwünschte Schwierigkeitsindizes

Welche Schwierigkeitsindizes sind in einem Test erwünscht? Die Frage läßt sich nur beantworten in Zusammenhang mit Trennschärfe und Homogenität. Hier darum nur eine vorläufige Antwort!

Schnelligkeitstests (Speed tests): Schnelligkeitstests sind solche Verfahren, bei denen die Durchführungszeit begrenzt ist. Die Aufgaben sollten leicht sein (p also hoch liegen). „Starke“ Probanden lösen viele, „schwache“ lösen wenige Items. Die Leistung wird gemessen durch die Zahl der Aufgaben, die in einer bestimmten Zeit gelöst werden.

Niveau-Tests (power tests): Bei reinen Niveau-Tests ist die Durchführungszeit nicht begrenzt. Ein Proband kann solange arbeiten, bis er die Items löst oder bis er aufgibt.

Bei dieser Klasse von Tests werden die Items nach aufsteigender Schwierigkeit angeordnet: Auf leichtere Items folgen schwerere. Im Idealfall gibt die laufende Nummer des Items, das der Proband als letztes zu bearbeiten vermag, das Leistungsniveau an, bis zu dem der Proband vordringt. Die Leistung wird allein durch die Zahl der gelösten Items gemessen.

Mischtests: Die meisten Tests fordern beide Anteile: Schnelligkeit und Niveauleistung. Darum werden die Items in aufsteigender Schwierigkeit angeordnet (die Schwierigkeiten bleiben in einem mittleren Bereich), und die Durchführungszeit wird begrenzt. Die Leistung wird gemessen durch die Zahl der in begrenzter Zeit gelösten Items.

4.2.1.4 Schwierigkeitsindex und andere Itemkennwerte

Der Schwierigkeitsindex „beeinflußt“ die anderen Itemkennwerte:

- Hat ein Item ein mittleres p (um 0.50), dann **ermöglicht** dies hohe Trennschärfen, verbürgt sie aber nicht.
- Streut p innerhalb eines Testes weit (z.B. von $p = 0.10$ bis $p = 0.90$), dann sinkt die Homogenität im Sinne von Interkorrelation.
Komplementär gilt: Variiert p innerhalb eines Testes in engen Grenzen (z.B. zwischen 0.20 und 0.30 oder zwischen 0.65 und 0.75), dann ermöglicht dies hohe Homogenität im Sinne von Interkorrelation (verbürgt sie aber nicht).
- Allein **von der Theorie** her betrachtet, wäre ein $p = 0.50$ der ideale Schwierigkeitsindex.

Hätten aber alle Items eines Tests den Wert $p = 0.50$ **und** wären auch bei allen Items jeweils dieselben Personen die Löser und dieselben Personen die Nichtlöser, dann zerfiele die Stichprobe immer in zwei Klassen: in Löser und Nichtlöser. Dieses Ergebnis widerspräche der Absicht, zwischen Probanden vielfältig zu differenzieren bzw. eine Skala zu erstellen, die

mehr als zwei Ausprägungen mißt.

Aus praktischem Interesse wählt man darum sowohl Items mit einem $p = 0.50$ als auch Items mit $0.50 > p < 0.50$: Man lockert die Homogenität auf zugunsten der Differenzierungsvielfalt.

- Schwierigkeitsindizes können mit den Stichproben wechseln, an denen sie erhoben werden. In diesem Phänomen zeigt sich die sogenannte **Stichprobenabhängigkeit** von Ergebnissen der klassischen Testtheorie.

4.2.2 Trennschärfe

Die Trennschärfe ist der wichtigste Itemkennwert, sie klärt die Position *eines* Items im Verband der anderen Items, indem sie einen Index liefert, der angibt, wie weit die ‚Menge der Löser‘ über alle Items hinweg identisch bleibt.

Diesen Index soll der **Vergleich mit einem Kriterium** erbringen. Dies kann ein äußeres oder ein inneres Kriterium sein.

- Beispiel für **ein äußeres Kriterium** sei das Urteil zweier Schreinermeister A und B über ihre Lehrlinge. Seien die ‚Items‘ Werkstücke, welche die Lehrlinge anfertigen.
 ⇒ Jeder Lehrling werde von Meister A danach eingestuft, wie ‚kunstfertig‘ er ist.
 ⇒ Das Werkstück jedes Lehrlings werde von Meister B danach bewertet, wie ‚kunstgerecht‘ es ist.
 Als trennscharf erweisen sich dann jene Werkstücke (Items), bei denen beide Urteilsreihen übereinstimmen.
- Beispiel für **ein inneres Kriterium** sei der Test-Score. Von jedem Probanden stehen zwei Werte zur Verfügung: sein Item-Score und sein Test-Score. Diese beiden Werte werden verglichen.

In der Regel wird die Trennschärfe an dem inneren Kriterium, dem Test-Score, ermittelt. **Vereinfacht** läßt sich darum die **Trennschärfe definieren** als die (biserielle) **Korrelation zwischen Item- und Test-Score**. Daher lautet das Kürzel für die Trennschärfe auch r_{it} : Korrelation (r) zwischen Item (i) und Test-Score (t)

Weil die Trennschärfe sich bestimmt durch Korrelation mit einem Kriterium, ist eine Ähnlichkeit zur **kriterienbezogenen Validität** gegeben (S. 98). Wir besprechen folgende Fragen:

- Berechnung der Trennschärfe (4.2.2.1),
- Teil-Ganz-Korrektur (4.2.2.2),
- Konvergente und diskriminante Trennschärfe (4.2.2.3),
- Trennschärfe und andere Itemkennwerte (4.2.2.4).

4.2.2.1 Berechnung der Trennschärfe

Berechnen läßt sich der Zusammenhang zwischen Einzelitem und Summenscore auf vielfältige Weise, z.B. durch

- biserielle Korrelation,
- Vier-Felder-Korrelation,
- Produkt-Moment-Korrelation,
- Kontingenzkoeffizient.

Hier sei die Ermittlung veranschaulicht an zwei Beispielen: an biserialer und Produkt-Moment-Korrelation.

Trennschärfe als punktbiserielle Korrelation

Für dichotome Item-Scores wird die Trennschärfe in der Regel als biserielle Korrelation bestimmt. Zur Verfügung stehen

- der **punktbiserielle** Korrelationskoeffizient für echt alternative Daten (z. B. Mann/Frau; schwanger/nicht schwanger) und
- der **biserielle** Korrelationskoeffizient für künstlich alternative Daten (z. B. Meßwerte ober-/unterhalb des Median).

Wir betrachten die Itemlösungen (0, 1) als echte Alternativen und berechnen die punktbiserielle Korrelation. Kasten 4.2-3 gibt eine Matrix vor.

Kasten 4.2-3:
Berechnung der Trennschärfe als punktbiserielle Korrelation

<div>- Zu den Termen: vgl. den Text zur Korrelationsformel (I).</div> <div>- + hinter Test-Score: Proband hat Item 1 richtig beantwortet.</div> <div>- - hinter Test-Score: Proband hat Item 1 falsch beantwortet.</div>								
Pbn	Items							
	1	2	3	4	5	6		
1	0	0	1	1	1	1	4	-
2	1	1	1	0	0	0	3	+
3	1	1	1	0	0	0	0	-
4	1	1	1	1	1	1	6	+
5	1	0	0	1	1	1	4	+
6	1	1	0	0	1	1	4	+
7	0	1	1	0	0	1	3	-
8	1	0	0	0	0	1	2	+
9	1	0	1	0	1	0	3	+
10	1	1	1	1	1	0	5	+
11	1	0	1	1	1	0	4	+
12	1	1	1	1	1	1	6	+
13	1	0	1	1	1	0	4	+

Berechnet sei die Trennschärfe für Item 1. Die Formel zur Berechnung der punktbiseriellen Korrelation (r_{pbis}) lautet:

$$(I) \quad r_{pbis} = \frac{M_r - M_x}{s_x} \sqrt{\frac{p}{q}}$$

Es bedeuten:

M_x = Mittelwert der Test-Scores X, hier: 3.69,

M_r = Mittelwert des Test-Scores der Probanden, die jenes Item **richtig** gelöst haben, dessen Trennschärfe berechnet wird, hier für Item 1: $(3 + 6 + 4 + \dots + 6 + 4)/10 = 4.10$,

s_x = Standardabweichung der Test-Scores X, hier: 1.69,

p = Schwierigkeitsindex des Items, hier für $p_1 = 0.77$,

$q = 1 - p$, hier für $q_1 = 0.23$.

Einsetzen für Item 1:

$$r_{pbis,1} = \frac{4.10 - 3.69}{1.69} \sqrt{\frac{0.77}{0.23}} = 0.44$$

Item 1 hat nach dieser Berechnung eine mittelhohe Trennschärfe von $r_{it} = 0.44$.

Trennschärfe als Produkt-Moment-Korrelation

Sind die Itemantworten nicht dichotom, sondern mehrfach abgestuft, so empfiehlt sich eine Berechnung, die dieser Vielfalt gerecht wird, zum Beispiel die Berechnung einer Produkt-Moment-Korrelation.

Die Produkt-Moment-Korrelation setzt voraus, daß die Item-Scores Intervallniveau erreichen und die Beziehung zwischen Item-Score und Test-Score linear ist.

Unter dieser Voraussetzung geben wir eine Zahlenmatrix vor in Kasten 4.2-4.

Kasten 4.2-4:
Berechnung der Trennschärfe als Produkt-Moment-Korrelation

- Zu den Termen: vgl. den Text zur Korrelationsformel (II)
- ΣX : Summe aller Scores von Item 1
- ΣY : Summe aller Test-Scores

Pbn	Items						Test-Score
	1	2	3	4	5	6	
1	4	4	3	4	5	4	24
2	3	2	2	3	2	1	13
3	1	2	1	2	1	2	9
4	5	4	5	3	4	5	26
5	4	3	3	3	5	3	21
6	3	4	2	1	5	4	19
1	1	2	1	2	2	1	9
8	1	0	0	0	0	1	2
9	3	1	3	2	2	1	12
10	4	5	4	3	3	4	23
11	4	3	3	4	5	4	23
12	5	4	3	4	5	3	24
13	3	2	3	3	3	2	16
$\Sigma X = 41$							$\Sigma Y = 221$

Für die Berechnung sei die Formel benutzt:

$$(II) \quad r = \frac{N \cdot \Sigma xy - \Sigma X \cdot \Sigma Y}{\sqrt{[N \Sigma X^2 - (\Sigma X)^2] [N \Sigma Y^2 - (\Sigma Y)^2]}}$$

Es bedeuten:

N = Zahl der Probanden, hier: 13

ΣX = Summe X (Item-Score), hierfür Item 1: $= 4 + 3 + \dots + 5 + 3 = 41$

ΣY = Summe Y (Test-Score), hier: $= 24 + 13 + \dots + 24 + 16 = 221$

ΣX^2 = Quadratsumme X, hier für Item 1: $= 4^2 + 3^2 + \dots + 5^2 + 3^2 = 153$

ΣY^2 = Quadratsumme Y (Test-Score), hier: $= 24^2 + 13^2 + \dots + 24^2 + 16^2 = 4423$

ΣXY = Produktsumme XY, hier: $= 4 \cdot 24 + 3 \cdot 13 + \dots + 5 \cdot 24 + 3 \cdot 16 = 814$

Einsetzen für Item 1:

$$r_{it,1} = \frac{13 \cdot 814 - 41 \cdot 221}{\sqrt{[13 \cdot 153 - 41^2] [13 \cdot 4423 - 221^2]}} = 0.93$$

Zwischen Item 1 und dem Test-Score ergibt sich eine Trennschärfe von $r_{it,1} = 0.93$, ein sehr hoher Wert (beruhend auf fiktiven Daten!).

4.2.2.2 Teil-Ganz-Korrektur

Wird die Trennschärfe berechnet, wie geschehen, geht jedes Item zweimal in die Berechnung ein: einmal als Item-Score und einmal als Teil des Test-Scores.

Beispiel aus Kasten 4.2-4: Der Test-Score des Pb 1 setzt sich zusammen aus seinen Item-Scores: $4+4+3+4+5+4=24$. Der **Test-Score 24** enthält demnach als Summanden **auch den Score 4 von Item 1**.

Um diese Selbstkorrelation zu eliminieren, zieht man den Item-Score jeweils von ‚seinem‘ Test-Score ab und korreliert den ‚korrigierten‘ Test-Score mit dem Item. Ein Beispiel gibt Kasten 4.2-5 mit den Werten von Kasten 4.2-4.

Kasten 4.2-5:
Berechnung der Trennschärfe: Teil-Ganz-Korrektur
(vgl. Text)

Die Teil-Ganz-Korrektur bezieht sich auf Item 1 . Die ‚Korrektur‘ besteht darin, daß vom Test-Score der Item-Score 1 abgezogen, dann eine neue Korrelation berechnet wird.					
-- ΣX : Summe der Item-Scores von Item 1.					
-- ΣY : Summe des korrigierten Test-Scores.					
	Item-Score	Test-Score		Test-Score	
Pbn	Item 1	<i>Unkorrigiert</i>	Korrektur	<i>Korrigiert</i>	
1	4	24	24 - 4	=	20
2	3	13	13 - 3	=	10
3	1	9	9 - 1	=	8
4	5	26	26 - 5	=	21
5	4	21	21 - 4	=	17
6	3	19	19 - 3	=	16
7	1	9	9 - 1	=	8
8	1	2	2 - 1	=	1
9	3	12	12 - 3	=	9
10	4	23	23 - 4	=	19
11	4	23	23 - 4	=	19
12	5	24	24 - 5	=	19
13	3	16	16 - 3	=	13
$\Sigma X =$ 41		$\Sigma Y =$ 180			

Für Item 1 ist einzusetzen:

$$r_{it, corr, 1} = \frac{13 \cdot 661 - 41 \cdot 180}{\sqrt{[13 \cdot 153 - 41^2] [13 \cdot 2948 - 180^2]}} = 0.89$$

Es ergibt sich ein Koeffizient von $r_{it, corr, 1} = 0.89$. Hier zeigt sich: Die Selbstkorrelation trägt in dem fiktiven Beispiel nur unerheblich zu dem (unkorrigierten) Wert von $r_{it, 1} = 0.93$ bei.

Beispiel: Den Effekt einer ‚Korrektur‘ veranschaulicht Kasten 4.2-6. Die Daten stammen aus einer Vorlage des Gießen-Tests. Beantwortet haben ihn 36 Studierende; es handelt sich demnach nur um ein **Demonstrationsbeispiel**.

Kasten 4.2-6:
Trennschärfe

<i>Unkorrigiert: r_{it} - Korrigiert: $r_{it,corr}$</i>			
<i>Gießen-Test</i>	Item	r_{it}	$r_{it,corr}$
Skala 1:	9	.49	.29
	16	.75	.58
Negative	23	.74	.55
versus	27	.20	.00
positive soziale	33	.77	.62
Resonanz	37	.67	.46

4.2.2.3 Konvergente und diskriminante Trennschärfe

Die Trennscharfe dient dazu, Items zu identifizieren, die alle (hoch) mit demselben Kriterium korrelieren. Dies soll sichern, daß alle Items dasselbe Merkmal erfassen. Ein Problem erwächst daraus, daß so gut wie kein Item nur ein einziges Merkmal erfaßt, vielmehr jedes auch andere Merkmalsanteile einschließt.

Beispiel: ‚Konzentration‘ werde gemessen mit Rechenaufgaben, etwa denen des ‚Konzentrations-Leistungs-Tests‘ (KLT, von Düker & Lienert, 1965). Rechenaufgaben jedoch provozieren nicht nur Konzentration, sondern auch andere Fähigkeiten, im KLT etwa auch ‚Rechenschnelligkeit‘, ‚Geschick im Umgang mit Zahlen‘, ‚Kenntnis von Rechenricks‘ (Sommer, 1973).

Ein Test-Autor sucht Items, die den Hauptbezug zur eigenen Skala einschließen. Wie weit dies zutrifft, darüber kann die Berechnung konvergenter und diskriminanter Trennschärfen Angaben machen.

Das Vorgehen besteht darin, den Item-Score zunächst mit seinem eigenen Test-Score zu korrelieren (konvergente Trennschärfe), dann aber auch mit den Test-Scores anderer Skalen, zu denen eine Verbindung vermutet wird (diskriminante Trennschärfe).

Beispiel: Ein Intelligenztest gliedere sich in mehrere Untertests. Die Items des Untertests 1 sollen vor allem mit dem Test-Score 1 hoch korrelieren, die Items des Untertests 2 mit dem Test-Score 2. Darüber hinaus lassen sich die Items des Untertests 1 mit dem Test-Score 2 korrelieren, die Items des Untertests 2 mit dem Test-Score 1. - Erwartet wird jetzt, daß bei den Items der Skala 1 die Korrelation zu Skala 1 höher ausfällt als die zu Skala 2 oder irgend einer anderen Skala: Die **konvergente** Trennschärfe soll höher ausfallen als die **diskriminante** Trennschärfe. -

Kasten 4.2-7 gibt ein **Demonstrationsbeispiel** (mit den Daten des Kastens 4.2-6).

Kasten 4.2-7:
Trennschärfen: Konvergent r_{ii} / Diskriminant $r_{it, dis}$

Gießen-Test	Item	r_{it}	$r_{it,dis}$: Skala 2-5				
			Fett: Koeffizient zu hoch				
		1	2	3	4	5	
Skala 1:	9	29	-08	01	-46	-42	
	16	58	09	-14	02	-29	
Negative	23	55	00	25	-47	-07	
versus	27	00	22	-02	-22	-09	
positive soziale	33	62	-26	10	-41	-14	
Resonanz	37	46	-06	-15	-01	-32	
<i>Koeffizienten der Kürze halber ohne Dezimalpunkt</i>							

Berechnet man konvergente und diskriminante Trennschärfen, dann ergibt sich ein Analogon zu einem faktorenanalytischen Ansatz. Die Matrix konvergenter und diskriminanter Trennschärfen läßt sich interpretieren analog zu einer Faktorenmatrix. Eine Übersicht solcher Art gibt beispielsweise der ‚Hamburger Persönlichkeitsfragebogen für Kinder (HAPEF-K)‘ (Wagner & Baumgärtel, 1978).

Kasten 4.2-8:
Faktorenmatrix:
Analogon zu einer Matrix konvergenter und diskriminanter Trennschärfen

Kursiv und fett:		Ladung auf eigenem Faktor I/eigener Skala 1					
Nichtkursiv:		Ladung auf fremden Faktoren/Skalen					
Fett und nichtkursiv:		Ladung zu hoch					
h^2 :		Kommunalität ²					
Gießen-Test	Item	Faktoren					h^2
		I	II	II	IV	V	
Skala 1:	9	-48	-32	02	14	05	36
	16	-05	-73	17	-24	12	64
Negative	23	-57	-23	18	05	47	64
versus	27	-35	23	44	-07	-13	40
positive soziale	33	-48	-44	-16	-03	42	63
Resonanz	37	-04	-76	00	-26	11	67
<i>Ladungen der Kürze halber ohne Dezimalpunkt</i>							

Umgekehrt kann darum auch eine Faktorenanalyse der Items die Dienste einer Berechnung von Trennschärfen übernehmen (vgl. die Matrix der Faktorenladungen im ‚Gießen-Test‘: Beckmann, Brähler & Richter, 1990, 117). Allerdings schließt die Itemanalyse einen Vorzug ein: Sie nötigt **dazu, im vorhinein** theoretisch festzulegen, welche Merkmale ein Test erfassen soll, welche Items darum zu einer (Sub-)Skala zusammengefaßt werden. Bei der (explorativen) Faktorenanalyse kann eine solche Zuordnung von Items und Skala offen blei-

2 *Kommunalität*: Summe der quadrierten Faktorenladungen. „Aufgeklärter“ Anteil der Gesamtvarianz je Variable (hier: je Item).

ben - es ist möglich, erst **im nachhinein** zu bestimmen, welche Items zu einer Skala gehören sollen und dann aus dem Inhalt der Skalen-Items das Merkmal zu ‚benennen‘, das die Skala erfaßt. - Kasten 4.2-8 gibt ein **Demonstrationsbeispiel** (erneut mit den Daten des Kastens 4.2-6).

Die Analogie besteht darin, daß beide Matrizen Zusammenhänge erkennen lassen zwischen **einer** Skala und **anderen** Skalen.

4.2.2.4 Trennscharfe und andere Itemkennwerte

Die Trennschärfe hängt sowohl mit dem Schwierigkeitsindex wie auch mit der Interkorrelation der Items zusammen.

- Was den **Schwierigkeitsindex** angeht, so ermöglicht, rein theoretisch, jeder Wert von p eine Trennschärfe von 1, ausgenommen die Werte $p=1$ und $p=0$. Aber nicht jede Kombination von p und r ist gleich effektiv. Am effizientesten im Sinne einer Differenzierung zwischen den Probanden ist die Kombination von $p = 0.50$ und $r_{it} = 1$, weil in diesem Falle die meisten Differenzierungen zwischen den Probanden möglich sind (Lienert, 1969, 126; Lienert & Raatz, 1994, 58).
- **Empirisch** ergibt sich ein Zusammenhang zwischen p und r_{it} , der sich als umgekehrtes U darstellt: Niedrigem oder hohem p entspricht ein niedriges r_{it} , mittlerem p ein hohes r_{it} .
- Was die **Interkorrelation** der Items angeht, so steigt die Trennschärfe, wenn Items hoch miteinander korrelieren; sie fällt, wenn Items niedrig miteinander korrelieren. Denn bei hoher Interkorrelation der Items ist es wahrscheinlicher, ja zwingend, daß, wer Item i löst, auch Item j löst und so einen hohen Test-Score erreicht. Bei niedriger Interkorrelation ist es weniger wahrscheinlich, daß, wer Item i löst, auch Item j löst und so einen hohen Test-Score erhält.

4.2.3 Homogenität

Die Items eines Tests sollen dasselbe Merkmal erfassen. Diesem Ziel setzt die Eigenart der Items eine Grenze. Items erfassen unterschiedliche Merkmalsfacetten. In diese Verschiedenheit gehen jedoch Schnittmengen gleicher (überlappender) Facetten mit ein. Das Maß für diese Übereinstimmung läßt sich als Homogenität bezeichnen.

Zwar hat sich in der klassischen Testtheorie keine einheitliche Auffassung des Begriffes Homogenität durchgesetzt. In allen Deutungen soll Homogenität aber den Grad angeben, in dem die Items eines Tests dieselbe Eigenschaft messen (Fischer, 1974, 127).

Es seien vier Konzepte besprochen:

- Homogenität im Sinne einer Interkorrelation (4.2.3.1),
- Homogenität im Sinne der Faktorenanalyse (4.2.3.2),
- Homogenität im Sinne einer Guttman-Skala (4.2.3.3),
- Homogenität im Sinne des Rasch-Modells (4.2.3.4).

Die zwei ersten Konzepte bleiben im Rahmen der klassischen Testtheorie, die zwei letzten überschreiten ihn (Fischer, 1974; Dieterich, 1973, 162).

4.2.3.1 Homogenität als Interkorrelation der Items

Die Trennschärfe eines Items schließt einen Bezug zum Gesamttest ein: den Vergleich aller Items mit demselben Kriterium. Die Homogenität als Interkorrelation der Items stellt ebenfalls einen Bezug zum Gesamttest her. Den Gesamtbezug repräsentiert aber nicht ein Kriterium, sondern der direkte Vergleich aller Items miteinander.

Das Ausmaß der Interkorrelation kann je nach Testziel variieren. Man spricht von homogenen und von heterogenen Tests:

- **Homogen** ist ein Test, dessen Items (vergleichsweise) hoch miteinander korrelieren. Inhaltlich bedeutet dies, daß die Items (nicht identische, aber) ähnliche Merkmalsfacetten repräsentieren.
- **Heterogen** ist ein Test, dessen Items (vergleichsweise) niedrig miteinander korrelieren. Inhaltlich besagt dies, daß die unterschiedlichen Items unterschiedliche Merkmalsfacetten erfassen.

Den **Grad** der Homogenität (H) gibt die Interkorrelation an. Ein Index läßt sich sowohl für jedes Item als auch für den Gesamttest ermitteln. (Der Homogenitätsindex für den Gesamttest nähert sich dem Charakter eines Reliabilitätskoeffizienten: vgl. S. 70.)

Kasten 4.2-9 veranschaulicht das Konzept an einem Beispiel. Die Einzelindizes erstrecken sich von $H = -0.062$ bis $H = 0.194$. Ebenso wie der Gesamtindex von $H = 0.091$ zeigen sie einen sehr heterogenen Test an.

Kasten 4.2-9:

Sechs Items eines (fiktiven) Tests werden interkorreliert. In der letzten Spalte erscheint der Homogenitätsindex jedes Items (H_{om}): Mittelwert der fünf ‚anderen‘ Itemkorrelationen (berechnet über Fishers z').

Unterhalb der Matrix ist ein Rechenbeispiel gegeben (H_1). Dort ist auch der Gesamtindex aufgeführt (H_{ges}): Mittelwert aller sechs Einzelindizes ($\Sigma H_{\text{om}}/6$). Beide Ergebnisse zeigen Heterogenität an.

Koeffizienten der Kürze halber ohne Dezimalpunkt!

Items							
Items	1	2	3	4	5	6	Horn
1	100	-13	-28	00	28	09	-008
2		100	68	01	03	-03	141
3			100	19	08	-20	122
4				100	58	-09	154
5					100	-08	194
6						100	-062
$H_1 = \frac{(-.13) + (-.28) + (.00) + (.28) + (.09)}{5} = -0.008$							
$H_{ges} = \frac{(-.008) + (.141) + (.122) + (.154) + (.194) + (-.062)}{6} = 0.091$							

Beziehung zu anderen Itemkennwerten

Homogenität als Interkorrelation hängt sowohl mit der Schwierigkeit als auch mit der Trennschärfe zusammen.

Was den **Schwierigkeitsindex** angeht, so gilt:

- Je mehr die Schwierigkeitsindizes streuen, desto niedriger korrelieren die Items miteinander, desto heterogener ist der Test.
- Je weniger die Schwierigkeitsindizes streuen, desto höher *kann* die Interkorrelation ausfallen.

Was die **Trennschärfe** angeht, so gilt:

- Hohe Interkorrelation der Items *ermöglicht* hohe Trennschärfen.
- Hohe Trennschärfen setzen hohe Interkorrelation der Items voraus.

4.2.3.2 Homogenität im Sinne der Faktorenanalyse

Items lassen sich (nicht nur interkorrelieren, sondern auch) faktorisieren. Homogen sind dann solche Items, die gemeinsam auf demselben Faktor (vergleichsweise) hoch laden. Unterschiedliche Faktoren mit den ihnen zugeordneten Items können unterschiedliche Skalen repräsentieren.

Diese Art von Homogenität ist der klassischen Testtheorie konform, sie läßt sich verstehen als ‚Verbesserung‘ der Homogenität im Sinne der Interkorrelation. - Überdies ergibt sich ein unmittelbarer Bezug: Die Faktorenladungen lassen sich deuten als Korrelation der Variablen/der Items mit einem Faktor. Darin läßt sich eine Analogie erkennen zur Trennschärfe als Korrelation der Items mit dem Test-Score.

4.2.3.3 Homogenität im Sinne einer Guttman-Skala

Der Guttman-Skalierung liegt ein Konzept von Homogenität zugrunde, welches das Modell der klassischen Testtheorie verläßt (Guttman, 1944).

Die Modellannahmen besagen: Ein Proband, der ein ‚schweres‘ Item löst, muß alle Items gelöst haben, die ‚leichter‘ sind. Das ‚schwerste Item‘, das ein Proband gelöst hat, zeigt den Ausprägungsgrad seiner Fähigkeit an. Wenn seine Fähigkeit ausreichte zur Lösung des schweren Items, muß sie erst recht ausreichen zur Lösung eines leichten Items.

Homogen ist ein Itemsatz, der diesen Modellannahmen entspricht. Es ergibt sich dann eine Matrix, die ein typisches ‚Dreieck der Lösungen‘ bildet, wie Kasten 4.2-10-A es anzeigt:

- In den **Spalten** wird das leichteste Item links, das schwierigste Item rechts plaziert. (Item a ist am leichtesten, Item f am schwierigsten.)
- In den **Zeilen** wird der Proband mit den meisten Lösungen oben, der Proband mit den wenigsten Lösungen unten plaziert. (Proband 1 hat sechs Items gelöst, Proband 6 nur ein Item.)

Kasten 4.2-10-A:
Homogenität im Sinne einer Guttman-Skala

+ Item gelöst; - Item nicht gelöst
Zum Verständnis siehe den laufenden Text!

Pbn	Items					
	a	b	c	d	e	f
1	+	+	+	+	+	+
2	+	+	+	+	+	—
3	+	+	+	+	—	—
4	+	+	+	—	—	—
5	+	+	—	—	—	—
6	+	—	—	—	—	—

Erläuterung zu Kasten 4.2-10-A: Das ‚Dreieck der Lösungen‘ verdeutlicht:

- **Item 1** ist am ‚leichtesten‘, weil es von allen Probanden gelöst wird.
- **Item 6** ist am ‚schwersten‘; nur *ein* Proband hat es gelöst.
- **Proband 1** ist am ‚tüchtigsten‘, weil er alle Items gelöst hat.
- **Proband 6** ist am ‚wenigsten tüchtig‘, er hat nur *ein* Item gelöst.

Ein Näherungsverfahren

Kasten 4.2-10-A veranschaulicht den *Idealfall* einer Guttman-Skala. Doch dürfte sich kein empirischer Datensatz finden, der diesem Ideal entspricht. Jane Loevinger hat eine Formel entwickelt, welche für empirische Datensätze die Annäherung an das ideale Muster ermittelt (1948; vgl. Lienert, 1969, 252):

$$H = \frac{N(\sum X^2 - \sum X) + \sum N_i^2 - (\sum X)^2}{2N(\sum(i \cdot N_i) - \sum X) + \sum N_i^2 - (\sum x)^2}$$

Es bedeuten:
H : Homogenitätsindex,
N : Zahl der Probanden,
X : Item-Score je Proband,
i : Rangplatz eines Items nach der Schwierigkeit p,
das leichteste Item erhält Rangplatz 1,
N_i : Anzahl der Probanden, die das Item mit dem Rangplatz i gelöst haben.

Kasten 4.2-10-B zeigt die leicht geänderte Matrix von Kasten 4.2-10-A: Bei Proband 3 weicht die Folge zu Item c, d und e von der Idealfolge ab. Für dieses Beispiel lautet das Ergebnis der Homogenität nach Loevingers Formel: H = 0.66.

$$H = \frac{6 \cdot (91 - 21) + 92 + 21^2}{2 \cdot 6 (59 - 21) + 92 - 21^2} = 0.66$$

Kasten 4.2-10-B:
Homogenität im Sinne einer Guttman-Skala
Angenäherte Matrix

(+ Item gelöst; - Item nicht gelöst) *Zum Verständnis siehe laufenden Text!*

Pbn	Items						X	X ²
	a	b	c	d	e	f		
1	+	+	+	+	+	+	6	36
2	+	+	+	+	+	+	5	25
3	+	+	—	+	+	—	4	16
4	+	+	+	—	—	—	3	9
5	+	+	—	—	—	—	2	4
6	+	—	—	—	—	—	1	1
p	1.00	0.83	0.50	0.50	0.50	0.33	21	91
i	1	2	4	4	4	6	21	
N _i	6	5	3	3	3	2	19	
i · N _i	6	10	12	12	12	12	59	
N _i ²	36	25	9	9	9	4	92	

Das Guttman-Modell ist ein **deterministisches** Modell, es beruht auf der Annahme:

- Eine Person **löst immer** ein Item, wenn ihre ‚Tüchtigkeit‘ ausreicht.
- Eine Person **löst nie** ein Item, wenn ihre ‚Tüchtigkeit‘ nicht ausreicht.

In diesem Punkte unterscheidet sich das Rasch-Modell.

4.2.3.4 Homogenität im Sinne des Rasch-Modells

Wie die Skalierung nach Guttman, so beruht auch eine Skalierung nach Rasch auf Ansätzen, die sich nicht decken mit den Annahmen der klassischen Testtheorie. Doch wird der **Zusammenhang zwischen ‚Fähigkeit‘ und ‚Itemlösung‘** nicht deterministisch (wie bei Guttman), sondern **probabilistisch** gefaßt.

Das besagt:

- **Die Annahme lautet nicht:** Jedes Mal, wenn die ‚Tüchtigkeit‘ eines Probanden größer ist als die ‚Schwierigkeit‘ eines Items, wird er das Item lösen.
- **Sondern sie lautet:** Es ist **wahrscheinlich**, daß ein Proband, dessen ‚Tüchtigkeit‘ größer ist als die ‚Schwierigkeit‘ eines Items, die Lösung findet.

Die Lösungswahrscheinlichkeiten lassen sich als Kurven darstellen. Als homogen gelten Rems, deren Verlaufskurven gleichartig sind und sich nur unterscheiden in ihrer Position auf der Fähigkeitsdimension. - Weitere Einzelheiten bietet die Skizze des Rasch-Modells (Kapitel 6, S. 151).

4.2.4 Testrevision und Itemselektion

Nach Abschluß der Itemanalyse ist zu entscheiden, welche Items zu behalten und welche zu eliminieren sind. Eine solche Entscheidung sollte nicht allein nach statistischen Gesichtspunkten getroffen werden, sondern inhaltliche Anliegen mitberücksichtigen.

Wir gehen in drei Schritten vor, wir besprechen:

- inhaltliche Fragen der Itemselektion (4.2.4.1),
- statistische Schritte der Itemselektion (4.2.4.2),
- weitere Gesichtspunkte einer Itemselektion (4.2.4.3).

4.2.4.1 Inhaltliche Fragen der Itemselektion

Inhaltliche Anliegen einer Itemselektion könnten sich in Fragen wie den folgenden artikulieren:

- Sind bestimmte Items zu behalten, um als *Eisbrecher* zu dienen?
- Sind Items zu behalten, weil sie einen *bestimmten Itemtyp* repräsentieren?
- Kann man auf bestimmte Items deswegen *verzichten*, weil *genügend andere Items* gleichen Inhalts vorliegen?
- Sind bestimmte Items (trotz ungünstiger Itemkennwerten) zu behalten, weil sie das *Testmerkmal* besonders *prägnant* repräsentieren?
- Sind bestimmte Items (trotz günstiger Itemkennwerten) deswegen zu *eliminieren*, weil sie sich *theoretisch-inhaltlich nicht so einordnen* lassen, wie es bei den Vorüberlegungen aussah?

- Sind Items zu *eliminieren*, weil sie *ethische Normen verletzen* (z.B. Verletzung der Intimsphäre, rassistische oder berufliche Diskriminierung)?
- Speziell bei *Persönlichkeitstests*:
 - ⇒ Sind die *Item-Antworten ausbalanciert*, d.h. gibt es etwa gleich viele Items, die mit ‚Ja/Stimmt‘, wie solche, die mit ‚Nein/Stimmt nicht‘ zu beantworten sind?
 - ⇒ Sind die Items frei von *Verzerrungstendenzen* (z.B. frei von sozialer Erwünschtheit)?

4.2.4.2 Statistische Schritte der Itemselektion

Es gibt unterschiedliche Prozeduren, welche die statistischen Entscheidungen einer Itemselektion regeln. Hier werden nur zwei besprochen:

- für *homogene* Tests: die Berechnung des Selektionskennwertes (A),
- für *heterogene* Tests: die Prüfung der Löserhäufigkeiten in den vier Quartilen der Test-Scores (B).

(A) Itemselektion bei *homogenen* Tests Berechnung des Selektionskennwertes

Homogene Tests setzen Schwierigkeitsindizes voraus, die in einem eng umschriebenen Intervall liegen (z. B. zwischen $p = 0.40$ bis $p = 0.60$). Damit solche Tests es aber erlauben, (nicht nur zwei oder drei, sondern) vielfältige Merkmalsabstufungen zu unterscheiden, sollten **auch** leichte und schwere Items erhalten bleiben.

Dem Zweck, eine Itemmenge mit einem breiten Band unterschiedlicher Schwierigkeiten auszuwählen, dient die Berechnung eines Selektionskennwertes (Lienert, 1969, 141-143; Lienert & Raatz, 1994, 117). Man kann das Ziel folgendermaßen umschreiben:

- Man will Items mit *niedriger Trennschärfe ausscheiden*, selbst wenn sie eine *mittlere (also günstige) Schwierigkeit* haben.
- Man will Items mit *hoher (also günstiger) Trennschärfe behalten*, selbst wenn sie eine *extreme (eine hohe oder niedrige) Schwierigkeit* haben.

Die Berechnung des Selektionskennwertes orientiert sich an Trennschärfe **und** Schwierigkeit, aber das wichtigere Kriterium bleibt die Trennschärfe.

$$Sel = \frac{r_{it}}{2 \cdot \sqrt{p \cdot q}}$$

Es bedeuten:

Sel : Selektionskennwert,

r_{it} : Trennschärfe,

p : Schwierigkeitsindex,
 q : $1 - p$.

Für die Itemselektion gilt die Regel: Man wähle jene Items, die einen höheren Selektionskennwert haben.

Um die Funktion zu veranschaulichen, sei in Kasten 4.2-11 die Trennschärfe konstant gehalten, in Kasten 4.2-12 dagegen der Schwierigkeitsindex.

In Kasten 4.2-11 wird die Trennschärfe konstant gehalten.

Kasten 4.2-11:

Funktion des Selektionskennwertes:

Die Trennschärfe bleibt konstant, die Schwierigkeit variiert von 0.10 bis 0.90.

Wie verhält sich der Selektionskennwert?

Trennschärfe konstant: $r_{it} = 0.60$	Schwierigkeit 0.10 - 0.90	Selektions- kennwert
	10	1.00
	.20	.75
	.30	.65
	.40	.61
	.50	.60
	.60	.61
	.70	.65
	.80	.75
	.90	1.00
Berechnungsbeispiel für $p = 0.70$: $Sel_{p=.70; r_{it}=.60} = \frac{0.60}{2 \cdot \sqrt{.70 \cdot .30}} = 0.65$		

Erläuterung zu Kasten 4.2-11: Die Trennschärfe wird konstant gehalten bei $r_{it} = .60$.

- Wir gehen aus von dem Item mit der Schwierigkeit $p = 0.50$; sein Selektionskennwert liegt bei $Sel = 0.60$.
- Wir betrachten die benachbarten Items mit Schwierigkeiten, die gegen $p = 0.10$ oder $p = 0.90$ gehen. Bleibt die **Trennschärfe** bei verschiedenen Items **gleich** (hier bei $r_{it} = 0.60$), dann steigt der Selektionskennwert um so höher, je weiter sich die Schwierigkeit von $p = 0.50$ entfernt.

Für die Itemselektion bedeutet das: Soll zwischen Items gleicher Trennschärfe gewählt werden, so „rät“ der Selektionskennwert zur Einbeziehung von schwierigen oder von leichten Items. *Diese Wahl „lockert“ die Homogenität auf:*

In Kasten 4.2-11 wird die Schwierigkeit konstant gehalten.

Kasten 4.2-12:**Funktion des Selektionskennwertes:**

Die **Schwierigkeit** bleibt konstant, die Trennschärfe variiert von 0.10 bis 0.90.

Wie verhält sich der Selektionskennwert?

Schwierigkeit konstant: $p = 0.60$	Trennschärfe 0.10 - 0.90	Selektionskennwert
	.10	.10
	.20	.20
	.30	.31
	.40	.41
	.50	.51
	.60	.61
	.70	.71
	.80	.82
	.90	.92
Berechnungsbeispiel für $r_{it} = 0.30$ $Sel_{r_{it}=0.30; p=0.60} = \frac{0.30}{2 \cdot \sqrt{.60 \cdot .40}} = 0.31$		

Erläuterung zu Kasten 4.2-12: Die Schwierigkeit wird konstant gehalten bei $p = 0.60$.

- Wir gehen aus von dem Item mit der Trennschärfe $r_{it} = 0.10$; sein Selektionskennwert liegt bei $Sel = 0.10$.
- Wir betrachten die Items mit Trennschärfen, die gegen $r_{it} = 0.90$ gehen. Bleibt die **Schwierigkeit** bei verschiedenen Items **gleich** (hier bei $p = 0.60$), dann steigt der Selektionskennwert mit steigender Trennschärfe.

Für die Itemselektion bedeutet das: Soll zwischen Items mit gleicher Schwierigkeit gewählt werden, so 'rät' der Selektionskennwert zur Einbeziehung jener Items, die eine höhere Trennschärfe haben.

Genereller Hinweis zum Selektionskennwert

„Bei der vom Vf. empfohlenen Technik . . . kann zwar die Aufgabenauswahl rein schematisch erfolgen, jedoch ist darüber hinaus eine Beachtung der Einzelkriterien stets wünschenswert“ (Lienert, 1969, 143).

(B)

Itemselektion bei heterogenen Tests:

Prüfung der Löser-Häufigkeiten in den vier Quartilen der Test-Scores

Bei heterogenen Tests empfehlen sich folgende Schritte der Itemselektion (Lienert, 1969, 144; Lienert & Raatz, 1994, 120):

- Ausgeschieden werden Items, deren *Schwierigkeit* über oder unter einer bestimmten Marke liegen, beispielsweise über $p = 0.85$ oder unter $p = 0.15$.

- Ebenso werden Items ausgeschieden, deren *Trennschärfe* unter eine bestimmte Marke fallen, etwa unter $r_{it} = 0.25$.
- Von den verbliebenen Items behält man jene mit höherer Trennschärfe.
- *Haben zwei oder mehrere Items die gleiche Trennschärfe, so behält man jenes Item, dessen Löser sich ,adäquat‘ über die vier Quartile der Test-Scores verteilen.*

Was ‚adäquat‘ hier besagt, sei verdeutlicht:

Bei zwei Items gleicher Trennschärfe bringt man die **Test-Scores** in eine Rangreihe, bildet das erste, zweite, dritte und vierte Quartil und prüft, wie sich die Item-Löser verteilen.

- Im ersten Quartil **erwartet** man: Wer einen niedrigen Test-Score hat (und deswegen zum 1. Quartil gehört), löst das Item nicht. Anders formuliert: Von den Item-Lösern sollten möglichst wenige ins 1. Quartil der Test-Scores fallen.
- Im vierten Quartil **erwartet** man: Wer einen hohen Test-Score hat (und deswegen zum 4. Quartil gehört), löst das Item. Anders formuliert: Von den Item-Lösern sollten möglichst viele zum 4. Quartil der Test-Scores gehören.
- Entsprechendes gilt vom zweiten und dritten Quartil.

Diese **Erwartung** beruht auf dem Konzept der Trennschärfe, wonach gilt: Die ‚Löser‘ eines Items sollen zu den Pbn mit **hohen** Test-Scores gehören (die Nicht-Löser zu den Pbn mit niedrigen Test-Scores).

Erwünscht sind demgemäß Items, deren **Löser** sich so verteilen, daß ‚wenige‘ zum ersten Quartil der Test-Scores (mehr zum zweiten und dritten Quartil) und ‚viele‘ zum vierten Quartil gehören. Diese Häufigkeitsverteilung der Löser sollte eine aufsteigende Gerade ergeben.

Beispiel: Gegeben seien Item 13 und 17, beide mit einem Schwierigkeitsindex von 0.50 und beide mit einer Trennschärfe von $r_{it} = 0.65$. Eines der beiden Items soll ausgeschieden werden. Die Trennschärfe gibt dafür in diesem Falle kein Kriterium ab. Um ein Kriterium zu gewinnen, ermitteln wir die Verteilung der Löser über die vier Quartile.

Beteiligt seien 200 Probanden. Ihre Test-Scores werden in eine Rangreihe gebracht. Es werden Quartile gebildet. Zu jedem Quartil gehören 50 Probanden. Zu klären ist, wieviele Probanden je Quartil zu den Lösern gehören.

Für das Beispiel gibt Kasten 4.2-13 die Verteilung an.

Kasten 4.2-13:**Itemselektion bei heterogenen Tests:**

Verteilung der Löser der Items 13 und 17 über die vier Quartile der Test-Scores
 Siehe den laufenden Text!

Quartil im Test-Score	Häufigkeit der Löser	
	Item 13	Item 17
1. 0– 6 Punkte	5	11
2. 7–14	18	12
3. 15–21	29	48
4. 22–30	48	29

Erläuterung zu Kasten 4.2-13:

- 5 Pbn, die **Item 13** lösen, gehören nach dem Test-Score ins 1. Quartil.
 (45 Pbn dieses Quartils sind Nicht-Löser.)
- 18 Pbn, die Item 13 lösen, gehören nach dem Test-Score ins 2. Quartil.
 (32 Pbn dieses Quartils sind Nicht-Löser.)
- 29 Pbn, die Item 13 lösen, gehören nach dem Test-Score ins 3. Quartil.
 (21 Pbn dieses Quartils sind Nicht-Löser.)
- 48 Pbn, die Item 13 lösen, gehören nach ihrem Test-Score ins 4. Quartil.
 (2 Pbn dieses Quartils sind Nicht-Löser.)

Welches Item ist in Kasten 4.2-13 zu behalten?

Zu behalten ist Item 13. Warum?

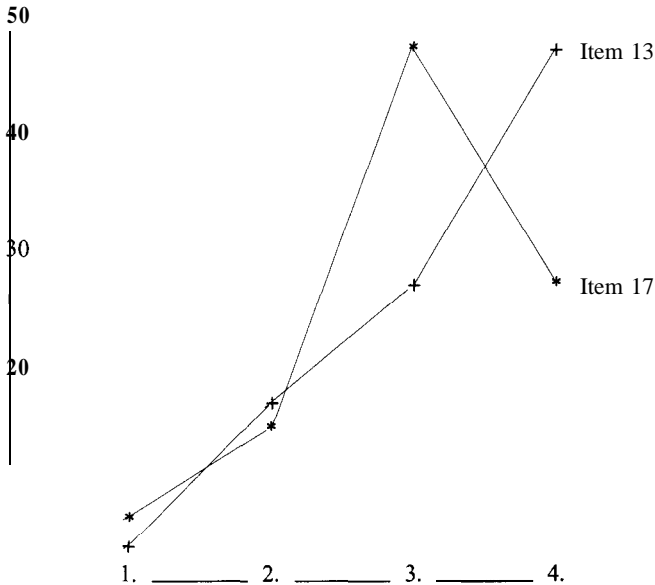
Bei **Item 13** verteilen sich die Löser gemäß dem Konzept der Trennschärfe in erwünschter Weise: Im 1. Quartil der Test-Scores finden sich ‚wenige‘, im 4. Quartil dagegen ‚viele‘ Löser. Die Verbindung der Häufigkeitspunkte ergibt fast eine Gerade.

Bei **Item 17** verteilen sich die Löser in unerwünschter Weise: im 1., 2. und im 3. Quartil finden sich ‚zu viele‘, im 4. Quartil ‚zu wenige‘ Löser. Die Verbindung der Häufigkeitspunkte ergibt keine Gerade.

Kasten 4.2-14 veranschaulicht die unterschiedliche Verteilung der Löser bei Item 13 und 17.

Kasten 4.2-14:**Itemselektion bei heterogenen Tests:**

Verteilung der Löser-Häufigkeit für Item 13 und 17 über die vier Quartile der Testscores
 Siehe Kasten 4.2-13 und den laufenden Text!

LÖSER-HÄUFIGKEIT**4.2.4.3 Weitere Gesichtspunkte einer Itemselektion**

Andere Gesichtspunkte, welche die Itemselektion mitbestimmen sollten, seien in Fragen gekleidet:

- Häufen sich gleiche Schwierigkeitsindizes in Bereichen, wo die Häufung unerwünscht ist? (Liegen z.B. zuviele Items in dem Bereich $p = .60$ bis $p = .80$?)
- Sind schwierige und leichte Items in gleichen Anteilen vertreten, so daß die Gesamtverteilung ausgewogen erscheint?
- Sind die Sprünge zwischen Items unterschiedlicher Schwierigkeit nicht zu groß, vor allem nicht, wenn sie in aufsteigender Reihe geordnet werden?

„Die simultane Selektion nach Trennschärfe und Schwierigkeit ist nicht einfach und endet nicht selten mit vielerlei Kompromissen, Man sollte auch immer diepsychologische Validität der einzelnen Aufgaben beachten, damit die Testendform auch den Laien in etwa befriedigt. Alle erfahrenen Testbearbeiter . . . sollten sich bei der Aufgabenselektion jedes Perfektionismus enthalten und ein Fingerspitzengefühl für diese Technik und die hierzu erforderliche Kompromißbereitschaft erwerben“ (Lienert, 1969, 139).

4.3 Ermittlung der Test-Gütekriterien

Vereinfacht gilt: Die Itemkennwerte charakterisieren den Test von seinen Einzelaufgaben, die Test-Gütekriterien von seiner Gesamtstruktur her. Unter drei Perspektiven wird bestimmt, wie angemessen der Gesamttest das empirische Relativ im numerischen Relativ abbildet.

Es geht um

- Standardisierung, genannt **Objektivität** (Kap. 4.3.1),
- Meßpräzision, genannt **Reliabilität** (Kap. 4.3.2),
- Merkmalssättigung, genannt **Validität** (Kap. 4.3.3).

HINWEIS: Die drei Hauptkriterien sind nicht disjunkt trennbar: *Objektivität läßt sich als Teilaspekt der Reliabilität betrachten, Validität als Sonderfall der Reliabilität, Reliabilität als Voraussetzung von Validität.*

Zu den drei Hauptgütekriterien kommen vier **Nebengütekriterien** hinzu (Lienert & Raatz, 1994, 7, 11-13):

- *Normierung:* Bezug zu einer Population,
- *Ökonomie:* Minimierung von Zeit- und Materialaufwand,
- *Nützlichkeit:* Bezug zur Praxis in Forschung oder Anwendung,
- *Vergleichbarkeit:* Bezug zu anderen Verfahren.

4.3.1 Objektivität

Objektivität bezeichnet das Maß, wie weit in der diagnostischen Situation eine **Standardisierung** des gesamten Testvorganges gelingt. Objektivität umfaßt „alle Variationsquellen, die zu Lasten unvollkommener Standardisierungen der einzelnen Phasen des diagnostischen Prozesses gehen“ (Michel & Conrad, 1982, 16).

Noch einmal ausführlicher: Objektivität gibt an,

- wie weit das Verhalten als empirisches Relativ *eindeutig quantifiziert* wird in Item- und Test-Scores als numerischem Relativ und
- wie weit diese Quantifizierung *sich eindeutig interpretieren* läßt.

Bei Durchführung, Registrierung und Auswertung desselben Tests soll das gleiche Verhalten eines Probanden immer in gleicher Weise quantifiziert und die quantifizierten Ergebnisse (die Test-Scores) immer in gleichem Sinne interpretiert werden. Objektivität bezeichnet demnach auch die **Unabhängigkeit der Testergebnisse vom Anwender**.

Bei einigen Autoren erhält Objektivität eine Sonderbedeutung. Sie bezeichnet die Undurchschaubarkeit eines Tests für den Probanden, betont also die Unabhängigkeit der Testergebnisse von Kognition oder Motivation der Probanden (Cattell, 1958; Fahrenberg, 1964; Häcker, 1982).

Wir gliedern den Stoff in die zwei Abschnitte: Arten der Objektivität (4.3.1.1) und *Probleme* der Objektivität (4.3.1.2).

4.3.1.1 Arten von Objektivität

Unterschieden werden in der Regel (mindestens) drei Arten:

- Durchführungs-,
- Auswertungs-
- und Interpretationsobjektivität.

Durchführungsobjektivität

Die Objektivität der Durchführung betrifft Raum und Zeit der diagnostischen Situation, die kognitiv-emotionale Verfassung des Probanden, darüber hinaus die Instruktion, welche die Testvorgabe und den Verlauf der Anwendung regelt.

Nur in seltenen Fällen läßt sich die **Durchführung** vollkommen standardisieren. Beispielsweise kann der Testleiter nur in begrenztem Maße die Befindlichkeit des Probanden vorhersehen und beeinflussen.

Eine approximative Standardisierung soll die **Testinstruktion** ermöglichen. Kasten 4.3-1 gibt ein Beispiel.

Kasten 4.3-1: Instruktion/Beispiele

Instruktion für **den Anwender**: Aus dem „Hamburg-Wechsler-Intelligenztest für Erwachsene (HAWIE)“ (Wechsler, 1964, 173):

„Bei der Durchführung des Tests muß der VL unbedingt die Anweisungen befolgen. Sie müssen wörtlich auswendig gelernt werden. Der VL soll die VP während des Tests nicht in ein Gespräch verwickeln; erlaubt sind nur notwendige Ermunterungen der VP. Die Anweisungen dürfen so oft wie erforderlich wiederholt, jedoch nicht erklärt werden.“

VL: Versuchsleiter / **VP:** Versuchsperson

Instruktion für **den Probanden**: Aus dem „Leistungsprüfsystem (LPS)“ (Horn, 1983, 7):

„In der folgenden zweistündigen Untersuchung soll festgestellt werden, welche Aufgabenarten dem Einzelnen leichtfallen, und was ihm weniger liegt. Die Zeit ist meist sehr kurz. Selbst der welcher ungewöhnlich schnell arbeitet, wird selten zur Lösung der schwersten Aufgaben kommen. Es ist jedoch wichtig, daß man sich immer Mühe gibt, damit man nicht falsch beurteilt wird...“

Jede Aufgabe wird in sehr einfachem Deutsch erklärt. Wer am Schluß der Erklärung noch nicht verstanden hat, was er machen soll, darf nicht laut fragen, sondern hebt nur seinen Arm. Es wird ihm dann nochmals persönlich erklärt werden, was er zu tun hat. Allerdings kann keinem später bei den schweren Aufgaben geholfen werden.”

Auswertungsobjektivität

Auswertungsobjektivität besteht darin, daß gleichen Itemantworten gleiche numerische Werte (Scores) zugeordnet werden.

- *Leicht* ist diese Standardisierung *bei gebundenen Items*, realisiert in Tests wie dem IST 70 von Amthauer (1973) oder dem LPS von Horn (1983).
- *Schwer* ist eine völlige Standardisierung *bei freien Items*. Hilfe bietet hier eine ausführliche Auswertungsanweisung, die viele Beispiele liefert. Um nur zwei Namen zu nennen: solche Auswertungshilfen geben Bäumler im Manual zum ‚Lern- und Gedächtnistest‘ (LGT 3: 1974), Schoppe im Manual zum ‚Verbalen Kreativitätstest‘ (VKT: 1975). - Kasten 4.3-2 führt ein Beispiel an.

Kasten 4.3-2:
Auswertungshilfen bei freien Items

Aus dem „Lern- und Gedächtnistest (LGT 3)“ (Bäumler, 1974, 36):		
In der 3.Aufgabe werden dem Probanden 20 Gegenstände gezeigt (bildlich), z.B. Kleiderbügel, Hammer, Roller.		
Zur Auswertung werden folgende Hilfen angeboten:		
Richtig	Noch gültig	Ungültig
<i>Kleiderbügel</i>	Aufhänger, Haken	<i>Bogen, Bumerang</i>
<i>Hammer</i>	Schlegel	<i>Werkzeug</i>
<i>Roller</i>	Zweirad	<i>Fahrrad, Rad</i>

Die Auswerterobjektivität läßt sich überprüfen, indem verschiedene Auswerter dasselbe Antwortprotokoll kodieren. Die Übereinstimmung kann korrelativ oder varianzanalytisch geschätzt werden.

Interpretationsobjektivität

Die Objektivität der Interpretation betrifft den Grad der Eindeutigkeit, mit der verschiedene Anwender dem gleichen numerischen Wert (dem Test-Score) die gleiche Merkmalsausprägung zuordnen. Wenn der Testautor Merkmalsbezeichnungen vorgibt und der Auswerter sie übernimmt, ist formal die Objektivität der Interpretation gegeben - es handelt sich um eine Sprachregelung.

Könnte eine Standardisierung der Interpretation nicht auch mehr anzielen: nämlich die Eindeutigkeit, die Experten als Interpreten einem Test-Score zuordnen? In dieser Deutung nähert sich das Konzept der Interpretationsobjektivität dem der Validität.

Für Interpretationshilfen gibt Kasten 4.3-3 ein Beispiel.

Kasten 4.3-3:
Festlegung der Interpretation/Beispiel

Im „Intelligenz-Struktur-Test 70 (IST 70)“ (Amthauer, 1973, 39) werden Interpretationen für alle Untertests angeboten. Hier zwei Beispiele:

„Was wird mit den Aufgabengruppen des I-S-T untersucht?

SE (Satzergänzung):

Urteilsbildung, common sense, Akzent im Konkret-Praktischen, Wirklichkeitssinn, Selbständigkeit im Denken.

WA (Wortauswahl):

Erfassen von sprachlichen Bedeutungsgehalten, Sprachgefühl, induktives sprachliches Denken, Einfühlungsfähigkeit. rezeptive Komponenten.“

Ein Interpretationsproblem: Ein Problem eigener Art kann sich daraus ergeben, daß zwei Probanden zwar den gleichen Score erreichen, der Score aber auf disjunkten Item-mustern beruht. Indiziert hier der gleiche Test-Score immer gleiches Merkmal und gleiche Ausprägung? - Rein formal gesehen: Ja! Items gelten als gleichwertig. Inhaltlich können Divergenzen auftreten.

Kasten 4.3-4 gibt ein Beispiel.

Kasten 4.3-4:
Interpretation: Gleicher Score/Divergierende Iteminhalte

Aus dem „Fragebogen zur Erfassung von Aggressivitätsfaktoren (FAF)“ (Hampel & Selg, 1975):

Proband A beantworte folgende drei Items in Schlüsselrichtung:

- 6: Es macht mir offen gestanden manchmal Spaß, andere zu quälen.
- 17: Als Kind habe ich manchmal ganz gerne andere gequält.
- 26: Mir hat es als Kind eigentlich Spaß gemacht, wenn andere von Eltern oder von Lehrern Prügel bezogen.

Proband B beantworte folgende drei Items in Schlüsselrichtung:

- 14: Zwischen anderen und mir gibt es oft Meinungsverschiedenheiten.
- 27: Ich hatte schon einmal solchen Zorn auf jemand, daß ich ihm den Tod wünschte.
- 40: Es macht mir Spaß, anderen Fehlern nachzuweisen.

Proband A scheint eher Aggression zu bejahen, die sich auf Verhalten, Proband B dagegen Aggression, die sich auf **Vorstellungen** bezieht. Beide erhalten den gleichen Test-Score: einen Rohwert von 3 Punkten.

Lassen sich beide Scores in dem Sinne gleicher Aggressivität interpretieren?

4.3.1.2 Probleme der Objektivität

Die Forderung nach voller Standardisierung begründet sich aus dem Anliegen, ein Merkmal **interaktionsfrei** zu messen. Paradox formuliert: Der Testleiter soll mit dem Probanden interagieren, ohne Interaktionseffekte hervorzurufen. Geht es damit aber nicht um Vorstellungen, die nicht einmal in der Physik zu realisieren sind, wie wir inzwischen wissen?

Eine **Gegenposition** bezieht die tiefenpsychologisch orientierte Therapie, indem sie Interaktionen mit dem Probanden gezielt in die Interpretation einbezieht, um so den diagnostischen und therapeutischen Prozeß zu verstehen und zu steuern. Rational verantwortbar bleibt eine solche Prozedur aber nur dann, wenn die subjektiven Effekte in Diskussion und **Supervision** (also in einem System rivalisierender Experten) kontrolliert werden.

Man darf aber nicht verkennen, daß in der **Objektivität ein zentrales Anliegen psychologischen Diagnostizierens** zur Sprache kommt: Der Untersucher will in einem Test jenen Anteil eines Merkmals erfassen, der dem Probanden zukommt (und nicht ihm selber, dem Anwender). Diese Probanden-Komponente findet der Untersucher nur heraus, wenn er seinen eigenen Anteil an der Messung isolieren und eliminieren kann oder wenn er zeigen kann, daß nicht er selber die Messung beeinflusst.

4.3.2 Reliabilität

Reliabilität charakterisiert das Meßinstrument Test unter dem Aspekt der Präzision. Implizit sind damit zwei Anteile angesprochen: wahrer Wert und Fehlerwert. Von beiden Anteilen her ist Reliabilität definiert worden:

- Reliabilität gilt als *Meßgenauigkeit* des Instrumentes unter Absehung vom Inhalt.
- Reliabilität gilt als Bestimmung des *Meßfehlers*, mit dem die Testwerte behaftet sind, unabhängig davon, für welchen Inhalt die Werte stehen.

Beide ‚Definitionen‘ beruhen auf Annahmen über den Zusammenhang zwischen wahrem Wert und Fehler, den sogenannten Axiomen der klassischen Testtheorie.

Wir besprechen fünf Problembereiche:

- Axiome der klassischen Testtheorie (4.3.2.1),
- Definition von Reliabilität (4.3.2.2),
- Veranschaulichung der Axiome und der Definition von Reliabilität (4.3.2.3),
- Modelle der Reliabilitätsberechnung (4.3.2.4),
- Test-Score und Vertrauensbereich (4.3.2.5),
- Kritische Differenzen (4.3.2.6).

4.3.2.1 Axiome der klassischen Testtheorie

über wahren Wert und Fehler wurden Annahmen formuliert, die sich selber nicht mehr ableiten lassen, aber ihrerseits Ableitungen begründen; diese Annahmen heißen Axiome. „Es sei . . . darauf hingewiesen, daß dieses axiomatische System zunächst formal-logische und somit nicht falsifizierbare Relationen zwischen definierten Modellkomponenten beschreibt. Die Brauchbarkeit

solcher syntaktischer Aussagen hängt jedoch im wesentlichen davon ab, ob sie die Realität der Meßvorgänge anhand psychologischer Tests hinreichend abdecken“ (Michel & Conrad, 1982, 19).

Hier seien die Axiome nur skizziert, nicht im einzelnen kommentiert (vgl. Fischer 1974; Lord & Novick, 1974; Michel & Conrad, 1982).

Axiom 1: Ein beobachteter Wert (X) setzt sich additiv zusammen aus wahren Wert (True score: T) und Fehlerwert (Error: E):

$$X = T + E$$

Der Fehler läßt sich weiter zerlegen (Magnusson, 1969, 114):

$$E = E_m + E_{adm} + E_g + E_{subj} + E_{Tf1}$$

Es bedeuten:

- E_m = E zufolge Erinnerung (memory),
- E_{adm} = E zufolge Anwendung (administration),
- E_g = E zufolge Rate-Effekten (guessing),
- E_{subj} = E zufolge subjektiver Auswertung,
also zufolge mangelnder Objektivität,
- E_{Tf1} = E zufolge fluktuierender wahrer Werte.

Axiom 2: Der Erwartungswert der Fehler [$a(E)$] und die Summe der Fehler [ΣE] sind gleich Null:

$$\epsilon(E) = \Sigma E = 0$$

Als Folgerung ergibt sich: Der Mittelwert der wahren Werte [$M-r$] ist gleich dem Mittelwert der beobachteten Werte [M_x]:

$$M_x = M_T$$

Voraussetzung: Axiom 2 setzt voraus, daß in „ $X = T + E$ “ je Individuum nur E variiert und T invariant bleibt. Mit anderen Worten.. Je Individuum gilt der wahre Wert als stabil. Fluktuierete (nicht nur der Fehler sondern auch) der individuelle wahre Wert, ließe sich die Variation nicht mehr eindeutig trennen in „wahre Anteile“ und „Fehleranteile“ (Faßnacht, 1995, 214).

Konsequenz: Die klassische Testtheorie läßt sich nur anwenden auf Merkmale, die stabil sind (auf sogenannte „traits“. Sie eignet sich nicht zur „Messung“ von Verhaltens-Prozessen.

Axiom 3: Fehlerwert und wahrer Wert korrelieren (p) nicht systematisch miteinander:

$$\rho_{T,E} = 0$$

HINWEIS: Aus den drei Axiomen lassen sich Aussagen über zwei Kovarianzen ableiten: über die Kovarianz von ‚wahren Wert und Fehlerwert‘

($cov_{T,E}$) und über die Kovarianz von ‚beobachtetem und wahren Wert‘ ($cov_{X,T}$). Es läßt sich zeigen, daß gilt:

$$cov_{T,E} = 0 \text{ und } cov_{X,T} = s_T^2$$

Diese Terme werden wir bei Besprechung der kriteriumsbezogenen Validität benötigen (S. 101).

Axiom 4: Wahrer Wert und Fehlerwert zweier verschiedener Tests (T_a , E_b) korrelieren (p) nicht systematisch miteinander:

$$\rho_{Ta, Eb} = 0$$

Axiom 5: Fehlerwerte zweier unterschiedlicher Tests (E_a , E_b) korrelieren (p) nicht systematisch miteinander:

$$\rho_{Ea, Eb} = 0$$

4.3.2.2 Definition von Reliabilität

Von den Axiomen leiten sich zwei Definitionen der Reliabilität ab:

- Reliabilität als *Quotient aus wahrer und beobachteter Varianz* oder
- Reliabilität als *Korrelation zweier Paralleltests*.

Beide Definitionen führen zu demselben Ergebnis. Hier sei nur die Bestimmung als Quotient eingeführt. Gegeben sind die drei Größen T, E, X (wahrer Wert, Fehlerwert und beobachteter Wert). Wiederholte Messungen führen zu drei Varianzen: s_T^2 , s_E^2 und s_X^2 .

Benötigt werden noch folgende Terme:

$r_{T,E}$: Korrelation zwischen wahren Wert und Fehler,

s_T : Standardabweichung der wahren Werte,

s_E : Standardabweichung der Fehler.

Die drei Varianzen (s_T^2 , s_E^2 , s_X^2) verhalten sich wie folgt:

$$s_X^2 = s_T^2 + s_E^2 + 2r_{T,E} \cdot s_T \cdot s_E$$

Nun gilt nach Axiom 3: $r_{T,E} = 0$. Somit entfällt der dritte Summand. Es bleibt:

$$s_X^2 = s_T^2 + s_E^2$$

Bei dieser Gleichung setzt die Definition von Reliabilität an. Als Kürzel für Reliabilität steht: r_{tt} . Es wird festgelegt:

$$r_{tt} = \frac{s_T^2}{s_X^2}$$

DEFINITION: Als *Reliabilität* gilt der *Quotient von ‚wahrer Varianz‘* (s_T^2) *und ‚beobachteter Varianz‘* (s_X^2), also der Anteil, den die wahre Varianz an der beobachteten Varianz erreicht.

Für die beobachtete Varianz s_X^2 (im Nenner) läßt sich einsetzen $s_T^2 + s_E^2$, so daß auch gilt:

$$r_{tt} = \frac{s_T^2}{s_T^2 + s_E^2}$$

Diese Festlegung ergibt eine *sinnvolle Definition*:

- Wenn die Fehlervarianz (s_E^2) gegen Null strebt, geht die Reliabilität gegen Eins.
- Wenn die Fehlervarianz gegen Unendlich strebt, geht die Reliabilität gegen Null (wegen $s_T^2 \geq 0$).

Zwei Varianten derselben Grundformel

Die Ausgangsformel läßt sich in zwei Varianten darstellen:

$$(I) \quad r_{tt} = \frac{s_T^2}{s_X^2}$$

$$\text{Nun gilt: } \begin{aligned} s_X^2 &= s_T^2 + s_E^2 \\ s_T^2 &= s_X^2 - s_E^2 \end{aligned}$$

Einsetzen im Zähler von (I) und Kürzen ergibt (II):

$$(II) \quad r_{tt} = 1 - \frac{s_E^2}{s_X^2}$$

HINWEIS: In der Formel (II) ist ein Fehlerterm enthalten (s_E^2), der dienen kann, einen sogenannten ‚Vertrauensbereich‘ zu berechnen, einen Bereich, in dem der wahre Wert liegt ($s_E^2 = s_X^2 [1 - r_{tt}]$, S. 90).

4.3.2.3 Veranschaulichung der Axiome und der Definition von Reliabilität

In Kasten 4.3-5 sei an einem Zahlenbeispiel veranschaulicht, wie sich eine beobachtete Varianz (s_X^2) in wahre Varianz und Fehlervarianz (s_T^2 , s_E^2) zerlegen läßt und wie die drei Größen in die Definition von Reliabilität eingehen.

Kasten 4.3-5:
Veranschaulichung der Axiome und der Definition von Reliabilität

<i>Pbn</i> :	Probanden	<i>cov_{TE}</i> :	Kovarianz von wahren Wert (T) und Fehler (E)		
<i>N</i> :	Zahl der Probanden	<i>cov_{XT}</i> :	Kovarianz von beobachtetem Wert (X) und wahren Wert (T)		
<i>M</i> :	Mittelwert				
<i>s²</i> :	Varianz				
<i>Spalte</i>	1	2	3	4	5
	Beobacht. W. <i>X</i>	Wahrer W. <i>T</i>	<i>Fehler</i> <i>E</i>	<i>Produkt</i> <i>T * E</i>	<i>Produkt</i> <i>X * T</i>
	4	3	+1	3	12
	3	4	-1	-4	12
	5	5	0	0	25
	7	7	0	0	49
	5	4	+1	4	20
	2	3	-1	-3	6
Σ	26	26	0	0	124
<i>M</i>	4.33	4.33	0	0	-
ΣX^2	128	124	4	-	-
<i>s²</i>	2.55	1.88	0.67	-	-
<i>s</i>	1.59	1.37	0.81	-	-
<i>cov_{TE}</i>		-	-	0	-
<i>cov_{XT}</i>	-	-	-	-	1.88

Erläuterungen zu Kasten 4.3-5:

Zu den Daten:

- Die *Daten* sind *fiktive* Werte, welche die drei Anteile veranschaulichen, die gemäß den Axiomen der klassischen Testtheorie in einem Meßwert angenommen werden: beobachteter, wahrer und fehlerhafter Wert (X, T und E). Die Zerlegung solcher Werte läßt sich empirisch weder verifizieren noch falsifizieren.

Zu den Axiomen:

- Zu **Axiom 1**: Jeder beobachtete Wert X zerlegt sich in die zwei Anteile T und E. **Beispiel (Spalte 1-3)**: Für Proband 1 zerlegt sich X = 4 in T = 3 und E= + 1.
- Zu **Axiom 2**: Die Summe der Fehler ist gleich Null, darum der Mittelwert der beobachteten Werte gleich dem Mittelwert der wahren Werte. **Beispiel (Spalte 3)**: $M_E = 0 \Rightarrow$ (Spalte 1 und 2): $M_X = 4.33 = M_T$
- Zu **Axiom 3**: Wahre Werte und Fehler korrelieren nicht miteinander: Ihre Kovarianz (*cov_{TE}*) ist gleich Null. Um zu demonstrieren, wie sich dieser Sachverhalt in den Daten darstellt, berechnen wir die Kovarianz nach der Formel:

$$\text{cov}_{TE} = \frac{\Sigma TE - \frac{\Sigma T \Sigma E}{N}}{N}$$

An der Formel wird ablesbar: Damit die Kovarianz der wahren Werte und der Fehler (cov_{TE}) gleich Null wird, muß - im Zähler - sowohl die Produktsumme der wahren Werte und der Fehler (ΣTE) als auch die Summe der Fehler (ΣE) den Wert Null annehmen.

Veranschaulichung:

1. Spalte 3 zeigt die Summe E ($\Sigma E = 0$), Spalte 4 die Produktsumme von T und E ($\Sigma TE = 0$). **Es folgt:** Die Kovarianz von wahren Wert und Fehler ist gleich Null: $\text{cov}_{TE} = 0$.
 2. Spalte 5 zeigt die Produktsumme von beobachteten und wahren Werten an ($\Sigma XT = 124$). **Es folgt:** Die Kovarianz von beobachteten und wahren Werten (Spalte 5: cov_{XT}) ist identisch mit der Varianz der wahren Werte (Spalte 2: s^2_T): $\text{cov}_{XT} = 1.88 = s^2_T$.
- Zu Axiom 4 und 5: Wahre Werte und Fehler verschiedener Test A und B korrelieren nicht systematisch zusammen. Zu diesen Axiomen bietet der Kasten 4.3-5 keine Veranschaulichung.

Zu den Varianzen und Kovarianzen:

Aus den Daten seien drei Varianzen und zwei Kovarianzen berechnet:

- Varianz der beobachteten Werte: $s^2_X = 2.55$
- Varianz der wahren Werte: $s^2_T = 1.88$
- Varianz der Fehlerwerte: $s^2_E = 0.67$
- Kovarianz von T und E: $\text{cov}_{TE} = 0.0$
- Kovarianz von X und T: $\text{cov}_{XT} = 1.88$

Zur Ableitung der Reliabilität aus den Axiomen:

- An dem Zahlenbeispiel sei der Zusammenhang zwischen beobachteter Varianz (s^2_X), wahrer Varianz (s^2_T) und Fehlervarianz (s^2_E) veranschaulicht:
 - (a) in einer Gleichung,
 - (b) in einem Diagramm.

Zu (a): Gleichung: $s^2_X = s^2_T + s^2_E$
 $2.55 = 1.88 + 0.67$

Zu (b): Diagramm: In dem Diagramm zerlegen sich die drei Varianzen in unterschiedliche Anteile.

$s^2_X = 2.55$	
$s^2_T = 1.88$	$s^2_E = 0.67$

- Nun sei aus den Varianzen (der fiktiven Daten) die Reliabilität geschätzt. Beide Berechnungen führen zu demselben Ergebnis.

⇒ zuerst nach Formel (I)

$$r_{tt} = \frac{s_T^2}{s_x^2} = \frac{1.88}{2.55} = 0.74$$

⇒ *dann nach Formel (II)*

$$r_{tt} = 1 - \frac{s_E^2}{s_x^2} = 1 - \frac{0.67}{2.55} = 0.74$$

4.3.2.5 Modelle der Reliabilitätsberechnung

Die Definition der Reliabilität und ihre Formalisierung in den beiden Formeln I und II ermöglichen es, die Meßgenauigkeit von Tests auf unterschiedlichen Wegen zu prüfen:

- In dem einen Falle werden eher Schätzungen der wahren Varianz,
- in dem anderen Falle eher Schätzungen der Fehlervarianz gesucht.

Von hier ergeben sich unterschiedliche Schätzmodelle:

- Retestreliabilität (I),
- Parallelttestreliabilität (II),
- Halbierungsreliabilität (III),
- Konsistenzschätzung (IV).

Item-sampling-Modelle der Reliabilitätsschätzung: *Die drei Modelle Parallelttestreliabilität, Halbierungsreliabilität und Konsistenzschätzung setzen ein anderes Schätzmodell voraus als die Retestreliabilität. Man spricht von Item-sampling-Modellen'. Es wird angenommen, daß ein Universum von Items existiert, die das gleiche Merkmal umschreiben. Aus diesem Universum werden parallele Itemstichproben gezogen.*

„Ungeachtet des geltenden testtheoretischen Modells beinhaltet der Begriff ‚Reliabilität‘ nicht ein einheitliches Konzept, sondern ist vielmehr ein Oberbegriff für eine Reihe von Konzepten, die jeweils nur bestimmte Aspekte der Meßgenauigkeit betreffen“ (Michel & Conrad, 1982, 38). - Die vier Modelle sind nicht äquivalent. Darum muß eine Mitteilung von Reliabilität auch die Art ihrer Ermittlung angeben.

(I) Retestreliabilität

Retestreliabilität besteht in der Genauigkeit, mit der bei denselben Probanden und mit demselben Test die Ergebnisse mehrerer Testurigen miteinander korrelieren. Demnach setzt die Retestreliabilität voraus:

- Denselben Probanden wird derselbe Test unter vergleichbaren Bedingungen wenigstens zweimal vorgegeben.

- Die Scores werden korreliert: die Höhe des Koeffizienten gilt als Schätzung der Meßgenauigkeit.

Es ergibt sich:

$$r_{tt} = \frac{s_T^2}{s_X^2} = \frac{\text{cov}_{1,2}}{s_1 s_2}$$

Es bedeuten:

- s_T^2 : Varianz der wahren Werte,
- s_X^2 : Varianz der beobachteten Werte,
- $\text{cov}_{1,2}$: Kovarianz der Werte von Testung I und Testung II,
- s_1, s_2 : Standardabweichung der Werte von Testung I und II.

Im Zähler steht zuerst die wahre Varianz (s_T^2), dann die Kovarianz der Werte von Testung I und Testung II ($\text{cov}_{1,2}$). Diese **Kovarianz** wird demnach als Schätzung der **wahren Varianz** gedeutet.

Im Nenner steht zuerst die beobachtete Varianz (s_X^2), dann die Standardabweichungen der Werte von Testung I und Testung II ($s_1 \times s_2$), ihr **Produkt** gilt demnach als Schätzung der **beobachteten Varianz**.

Einflüsse auf die Retestreliabilität

Auf die Wiederholung des Tests können Einflüsse einwirken, die in der zweimaligen Testung nicht vorgesehen sind: vor allem Zeitabstand zwischen zwei Testungen, Gedächtniseffekte, generell jede Art von Lernen, aber auch Merkmalsfluktuation.

Retestreliabilität und Stabilität des Ziel-Merkmals

Das Paradigma der Retestreliabilität beruht auf einer fundamentalen Voraussetzung: Das Zielmerkmal muß relativ stabil bleiben. Nur dann ist Wiederholung von Testungen ein Weg, Fehlervarianz und wahre Varianz zu schätzen.

Dies sei veranschaulicht: Zwei Testungen mögen zwei gleiche Scores liefern, etwa zweimal den Wert 107. Der wahre Wert liege in Testung I bei 105, in Testung II dagegen bei 95; der Fehler betrüge für I demnach 2, für II dagegen 12. Der wahre Wert, als Indikator des Ziel-Merkmals, wurde von Test zu Retest also ‚fluktuieren‘. Dann aber wurde der gleiche Wert von 107 gleiche Merkmalsausprägung **vortäuschen**. - Demnach gilt umgekehrt: Nur wenn der gleiche Test-Score auf gleichen wahren Werten (auf einem stabilen Zielmerkmal) beruht, ist der Schluß auf Meßpräzision berechtigt.

Stabilität im weiteren Sinne: Wegen des Zusammenhanges mit der Stabilität des Merkmals wird Retestreliabilität oft auch als **Stabilität** bezeichnet, sie umschließt in diesem Falle:

die Genauigkeit des Instrumentes
 UND die Varianz aus Fehlern des Instrumentes
 UND die Varianz des Merkmals.

Stabilität im engeren Sinne: Zuweilen wird Stabilität aber auch in einem engeren Sinne gefaßt, die Varianz des Merkmals wird ausgeschlossen. Dann bezeichnet Stabilität:

die Genauigkeit des Instrumentes
 UND die Varianz aus Fehlern des Instrumentes
 OHNE die Varianz des Merkmals.

In diesem zweiten Falle liegt die Stabilität höher als die Retestreliabilität, setzt aber voraus, daß die Fluktuation des Merkmals anderswoher bekannt ist und eliminiert werden kann.

Unterschiedliche Itemmuster als Problem; der Retestreliabilität

Ein weiteres Problem kann sich daraus ergeben, daß bei Testung I und II zwar die Scores gleich hoch ausfallen, aber auf disjunkten Itemmengen beruhen, bei Testung I etwa auf den Items 1 3 5 7 9, bei Testung II auf den Items 2 4 6 8 10. Kann hier der gleiche Test-Score als Indikator für gleiche Verhaltensdisposition dienen?

Mit dem letzten Satz mündet die Frage der Reliabilität in eine Frage nach der Validität der Items. - Das Problem ergab sich schon bei Diskussion der Objektivität (S. 69). Es sei hier nur in Erinnerung gerufen.

(II) **Paralleltestreliabilität**

Paralleltestreliabilität wird bestimmt als Korrelation zwischen Test A und seinem Paralleltest B bei denselben Probanden. Demnach wird derselben Stichprobe ein Test A vorgelegt (gegebenenfalls auch C, D, E). Die Übereinstimmung der Test-Scores, ermittelt als Korrelation, gilt als Indikator für Meßgenauigkeit.

Als wichtigste Voraussetzung geht in diese Schätzung ein, daß Test A und Test B äquivalent sind. Dies schließt ein, daß die wichtigsten Kennwerte in beiden Tests gleich sind; dies betrifft zum mindesten

- die Verteilungskennwerte (Mittelwerte, Varianzen, Kovarianzen bei mehr als zwei Tests),
- die Reliabilitäten (je für sich) und
- die Validitäten (je für sich).

Wilks hat eine Prüfgröße entwickelt, genannt Lambda, die zu bestimmen erlaubt, wie weit die zentralen Kennwerte paralleler Testentwürfe äquivalent sind (Wilks, 1946; Dieterich, 1973, 153; Lienert & Raatz, 1994, 30 1-306).

„Leider wird meist allzu sorglos verfahren, indem man annähernd gleichartige Tests miteinander korreliert und das Resultat als Reliabilität interpretiert“ (Fischer, 1974, 39).

Für die Korrelation gilt:

$$r_{tt} = \frac{s_T^2}{s_x^2} = \frac{\text{cov}_{a,b}}{s_a \cdot s_b}$$

Es bedeuten:

s_T^2 : Varianz der wahren Werte,

s_x^2 : Varianz der beobachteten Werte,

$\text{cov}_{a,b}$: Kovarianz von Test A und Paralleltest B,

s_a, s_b : Standardabweichung von Test A und Paralleltest B.

Geht man wieder von der Definition der Reliabilität als Quotient von wahrer und beobachteter Varianz aus, dann folgt: Analog zur Retestreliabilität gilt die Kovarianz zwischen Test A und Test B ($\text{cov}_{a,b}$) als Indikator der wahren Varianz, das Produkt der Standardabweichungen von Test A und von Test B als Indikator der beobachteten Varianz ($s_a \cdot s_b$).

Probleme der Paralleltestreliabilität

Wer Test A kennt (weil er ihn bearbeitet hat), kann die Lösungsprinzipien auf Test B übertragen. Demnach ist mit ähnlichen Lerneffekten zu rechnen wie bei einem Retest.

Ebenso: Die beiden Itemstichproben sollen gleich, aber nicht identisch sein (Horst, 1971, 313, 352). Es soll dasselbe Merkmal gemessen werden, aber nicht mit denselben Items.

(III) Halbierungsreliabilität

Die Halbierungsreliabilität besteht, vereinfacht gesagt, in der Korrelation zwischen zwei Hälften desselben Tests bei denselben Probanden.

Ein Test wird nur einmal vorgelegt. Auf unterschiedlichen Wegen wird die Gesamtmenge der Items in zwei Hälften aufgeteilt. Die Aufteilung ist so vorzunehmen, daß die beiden Hälften ähnlich strukturiert sind wie zwei Paralleltests.

Je Proband wird dann für beide Hälften ein Test-Score gebildet. Diese werden korreliert. Die Höhe der Übereinstimmung gilt als Indikator der Meßgenauigkeit - die durch Hochrechnung korrigiert werden kann.

Als Möglichkeiten, beide Hälften zu vergleichen, bieten sich an:

- Die ‚untere Hälfte‘ des Itemsatzes wird verglichen mit der ‚oberen Hälfte‘, beispielsweise die Items 1-11 mit den Items 12-22.
- Items, die *ungerade* Nummern haben, werden verglichen mit Items, die *gerade* Nummern haben, beispielsweise die Items 1 3 5 7 mit den Items 2 4 6 8 .
- Der Gesamtsatz der Items wird *nach Zufall* in zwei Hälften eingeteilt, diese dann (wenn sie parallel ausfallen) verglichen.

Berechnung und Hochrechnung der Halbierungsreliabilität

Bei der Halbierungsreliabilität wird zunächst die Korrelation der beiden Hälften ermittelt, in der Regel als Produkt-Moment-Korrelation. Dann wird der ermittelte Koeffizient ‚korrigiert‘, d.h. hochgerechnet nach der sogenannten **Spearman-Brown-Formula** (prophecy formula). Zunächst sei ein Beispiel, dann die Hochrechnung, schließlich eine Begründung der Hochrechnung gegeben:

- Kasten 4.3-6 bringt ein *Zahlenbeispiel* zur Berechnung der Halbierungsreliabilität.
- Unter dem Kasten 4.3-6 folgen die *Hochrechnung* und
- ihre *Begründung*.

Kasten 4.3-6 veranschaulicht an einem *Zahlenbeispiel* die Berechnung der Halbierungsreliabilität.

Kasten 4.3-6:
Halbierungsreliabilität

Die Hälften werden gebildet nach ‚geraden/ungeraden‘ Items:								
H1: Summe aus Item 1, 3 und 5								
H2: Summe aus Item 2, 4 und 6								
Pbn	Items						H1	H2
	1	2	3	4	5	6		
1	2	4	3	3	4	4	9	11
2	3	4	5	2	4	4	12	10
3	5	6	4	5	6	2	15	13
4	4	3	4	2	4	3	12	8
5	2	1	1	2	4	1	1	4
6	4	5	6	5	5	2	15	12

Die Korrelation zwischen Hälfte 1 (H1) und Hälfte 2 (H2) in Kasten 4.3-6 beträgt:

$$r_{1,2} = 0.79$$

Diese Korrelation wird ‚hochgerechnet‘ nach der Spearman-Brown-Formula, die für den Fall der Halbierung lautet:

$$r_{uc} = \frac{2 \cdot r_{1,2}}{1 + r_{1,2}}$$

Es bedeuten:

r_{ttc} : Reliabilität korrigiert nach Spearman-Brown,

$r_{1,2}$: Korrelation der beiden Testhälften.

Einsetzen und Berechnen ergibt:

$$r_{ttc} = \frac{2 \cdot 0.79}{1 + 0.79} = \mathbf{0.88}$$

Der unkorrigierte Koeffizient, die Korrelation beider Hälften, beträgt $r_{tt} = 0.79$. Hochgerechnet ergibt sich eine korrigierte Halbierungsreliabilität von $r_{ttc} = 0.88$. Der erste Koeffizient liegt gerade unterhalb der für Reliabilität wünschenswerten Höhe (von $r = 0.80$), der zweite überschreitet sie erheblich.

Zur Ableitung der Hochrechnungsformel: der Spearman-Brown-Formula

Die Ableitung der Spearman-Brown-Formula geht aus von der Definition der Reliabilität: $r_{tt} = s^2_T / s^2_X$. Es läßt sich zeigen: Bei Verlängerung (oder Verkürzung) eines Tests ändern sich Zähler (s^2_T) und Nenner (s^2_X) in unterschiedlicher Weise.

Bei **Verlängerung** gilt: Die wahre Varianz (im Zähler) wächst rascher als die beobachtete Varianz (im Nenner). Das heißt, der Anteil der wahren Varianz an der beobachteten Varianz nimmt zu. Somit wächst (der Betrag des gesamten Quotienten, also) die Reliabilität.

Dies sei für den Fall der **Verdoppelung** eines Itemsatzes **demonstriert**. Zähler (s^2_T) und Nenner (s^2_X) werden getrennt betrachtet.

ZÄHLER

Für den Zähler gilt bei Verdoppelung, daß die **wahren** Varianzen von Teil 1 (bisheriger Itemsatz) und von Teil 2 (‘neuer‘ Itemsatz) sich addieren wie folgt:

$$(1) \quad s_{T(2)}^2 = s_{T1}^2 + s_{T2}^2 + 2 \cdot r_{T,12} \cdot s_{T1} \cdot s_{T2}$$

Es bedeuten:

$s_{T(2)}^2$: Summe der wahren Varianzen von Teil 1 und Teil 2,

s_{T1}^2, s_{T2}^2 : Wahre Varianz von Teil 1 und von Teil 2,

s_{T1}, s_{T2} : Standardabweichung der wahren Werte von Teil 1 und von Teil 2,

$r_{T,12}$: Korrelation zwischen den wahren Werten von Teil 1 und Teil 2.

Da diese Varianzen fehlerfrei sind (per definitionem!), gilt:

$$s_{T1}^2 = s_{T2}^2 = s_T^2$$

$$s_{T1} = s_{T2} = s_T$$

$$r_{T,12} = 1$$

Somit läßt sich (I) auch schreiben:

$$(II) \quad s_{T(2)}^2 = s_T^2 + s_T^2 + 2 \cdot 1 \cdot s_T \cdot s_T \\ = 2s_T^2 + 2s_T^2$$

$$(III) \quad s_{T(2)}^2 = 4s_T^2$$

NENNER

$$(IV) \quad s_{x(2)}^2 = s_{x1}^2 + s_{x2}^2 + 2 \cdot r_{x,12} \cdot s_{x1} \cdot s_{x2}$$

Es bedeuten:

$s_{x(2)}^2$: Summe der **beobachteten** Varianzen von Teil 1 und Teil 2,

s_{x1}^2, s_{x2}^2 : beobachtete Varianz von Teil 1 und von Teil 2,

s_{T1}, s_{T2} : Standardabweichung der beobachteten Werte von Teil 1 und von Teil 2,

$r_{x,12}$: Korrelation zwischen den beobachteten Werten von Teil 1 und Teil 2.

Da angenommen wird, daß die beiden Teile parallel sind, gilt:

$$s_{x1}^2 = s_{x2}^2 = s_x^2$$

$$s_{x1} = s_{x2} = s_x$$

Im Unterschied zum Zähler gilt nicht, daß die Korrelation zwischen den beiden Hälften den Wert 1 erreicht, und zwar wegen der Fehleranteile in den beobachteten Varianzen. Demnach gilt:

$$r_{x,12} < 1$$

Nun läßt sich (IV) auch schreiben:

$$(V) \quad s_{x(2)}^2 = s_x^2 + s_x^2 + 2 \cdot r_{12} \cdot s_x \cdot s_x \\ = 2s_x^2 + 2 \cdot r_{12} \cdot s_x^2$$

$$(VI) \quad s_{x(2)}^2 = 2s_x^2 (1 + r_{x,12})$$

Einsetzen von (III) und (VI) in die Reliabilitätsformel ergibt:

$$(VII) \quad r_{ucorr} = \frac{(III)}{(VI)} = \frac{4s_T^2}{2s_x^2(1+r_{x,12})} = \frac{2}{s_x^2} \frac{s_T^2}{(1+r_{12})}$$

Der Quotient $\frac{s_T^2}{s_x^2}$ ist die Definition der Reliabilität.

Anstelle dieses Quotienten kann demnach auch r_{tt} stehen. Im Beispiel der Halbierungsreliabilität wird r_{tt} ermittelt als Korrelation zwischen Teil 1 und Teil 2, also als r_{12} . Somit läßt sich VII auch schreiben:

$$r_{ttcorr} = \frac{2 r_{tt}}{1 + r_{12}} = \frac{2 r_{12}}{1 + r_{12}}$$

Das aber ist die Spearman-Brown-Formula für den Fall der Verdoppelung eines Itemsatzes.

Verallgemeinerung der Korrekturformel

Ein Test läßt sich nicht nur halbieren, sondern ebenso dritteln oder vierteln usw. Für diesen allgemeinen Fall lautet die Spearman-Brown-Formula:

$$r_{uc} = \frac{n \cdot r_{tt}}{1 + (n - 1) \cdot r_{tt}}$$

Neu ist der Term ,n':

$$n = \frac{\text{Itemzahl „nach“ Korrektur}}{\text{Itemzahl „vor“ Korrektur}}$$

Die allgemeine Formel wird dazu verwandt, vorauszuschätzen, wie hoch die Reliabilität ausfallen wird, wenn ein Test um eine bestimmte Anzahl von Items verlängert oder verkürzt wird, aber auch, um zu schätzen, wieviele Items man einem Test hinzufügen muß, um eine angezielte Reliabilitätshöhe zu erreichen.

Kasten 4.3-7 bringt zwei Beispiele für die Anwendung der Spearman-Brown-Formula.

Kasten 4.3-7:

Anwendung der Spearman-Brown-Formula: Zwei Beispiele

1. Um wieviel **wächst** die Reliabilität, wenn ein Test mit $r_{tt} = 0.60$ von 40 Items auf 47 Items verlängert wird?

Antwort: Die Reliabilität wächst auf $r_{tt} = 0.64$.

Weg: In der Spearman-Brown-Formula ist zunächst ,n' zu ermitteln, dann einzusetzen. Die Zahl der Items nach der Korrektur (Zähler) beträgt 47, die Zahl *vor* der Korrektur (Nenner) dagegen 40.

Einsetzen:

$$n = 47/40 = 1.175$$

$$r_{ttcorr} = \frac{1,175 \cdot 0.60}{1 + (1,175 - 1) \cdot 0.60} = 0.64$$

Die Reliabilität wächst auf $r_{tt} = 0.64$.

2. Ein Test besteht aus 18 Items und hat eine Reliabilität von $r_{tt} = 0.75$. Wieviel Items müssen hinzukommen, damit die Reliabilität auf $r_{ttc} = 0.87$ ansteigt?

Antwort: Dem Test müssen 22 Items hinzugefügt werden, so daß er 40 Items zählt.

Weg: Gegeben sind zwei Werte: r_{ttc} und r_{tt} . Unbekannt ist n. *Die Formula ist aufzulösen nach n:*

$$n = \frac{r_{tt} \cdot (r_{tt} - 1)}{r_{tt} \cdot (r_{tt} - 1)}$$

Einsetzen:

$$n = \frac{0.87 \cdot (0.75 - 1)}{0.75 \cdot (0.87 - 1)}$$

Also gilt:

$$n = \frac{\text{Itemzahl „nach“ Korrektur } (n_{\text{nach}})}{\text{Itemzahl „vor“ Korrektur } (n_{\text{vor}})}$$

Bekannt sind demnach: $n = 2.33$ und $n_{\text{vor}} = 18$.

Unbekannt ist n_{nach} . Es gilt die Relation:

$$n_{\text{nach}} = n \cdot n_{\text{vor}}$$

Einsetzen: $n_{\text{nach}} = 2.33 \cdot 18 = 40.15$

Die Itemzahl „nach“ der Korrektur beträgt demnach 40.

Halbierungsreliabilität bei ungleichen Hälften

Die Spearman-Brown-Formel anzuwenden ist nur berechtigt, wenn verlängerte und verlängerte Teile äquivalent sind. Ist die Äquivalenz nicht gegeben, dann ist eine Korrektur nach Spearman-Brown unberechtigt,

Um Ungleichheiten der Teile zu berücksichtigen, wurde die Spearman-Brown-Formel ergänzt, so von Flanagan, von Kuder-Richardson, von Cronbach, von Kristof (vgl. Dieterich, 1973, 155-158; Kranz, 1981, 202-206; Lienert & Raatz, 1994, 185-191; Lord & Novick, 1974, 82-98). Kasten 4.3-8 gibt zwei Beispiele solcher Alternativformeln.

Kasten 4.3-8:

Alternativformeln zur klassischen Spearman-Brown-Formel: Zwei Beispiele

Nach Kristof

$$r_{tt} = \frac{2}{N-1} + \frac{N-3}{N-1} \cdot \frac{4 s_1 s_2 r_{12}}{s_1^2 + s_2^2 + 2 s_1 s_2 r_{12}}$$

Es bedeuten:

- N : Anzahl der Probanden,
- s_1^2, s_2^2 : Varianz von Teil 1 und Teil 2,
- $s_1 s_2$: Standardabweichung von Teil 1 und Teil 2,
- r_{12} : Korrelation zwischen Teil 1 und Teil 2.

Nach Cronbach: α -Koeffizient

$$\alpha = \frac{n}{n-1} \cdot \frac{s_1^2 - \sum s_{item}^2}{s_1^2}$$

Es bedeuten:

- n : Anzahl der Items,
- s_1^2 : Varianz des Gesamttests,
- s_{item}^2 : Varianz der Items.

EXKURS: Generalisierbarkeitstheorie als Erweiterung des klassischen Reliabilitätskonzeptes

Vom Konzept der Spearman-Brown-Korrektur her läßt sich auf eine Erweiterung des klassischen Reliabilitätskonzeptes verweisen. Das Gemeinsame liegt darin, daß erhobene Werte dazu dienen, ‚neue‘ Werte vorzuschätzen.

Erweitert wurde das klassische Konzept der Reliabilität durch die Generalisierbarkeitstheorie von Cronbach, Rajaratnam und Gleser (1963; vgl. Cronbach, Gleser, Nanda & Rajaratnam, 1972; Kamp, 1976). Eine knappe, sehr übersichtliche Einführung bietet beispielsweise Nußbaum bei Klauer (1987).

Die Generalisierbarkeitstheorie fordert, daß, bevor ein Konstrukt gemessen wird, ein ‚Universum zulässiger Beobachtungen‘ definiert werde. Festgelegt wird, unter welchen ‚Facetten‘ und unter welchen ‚Bedingungen‘ das Konstrukt beobachtet werden solle. Die Begriffe seien kurz erläutert.

Das Design der Generalisierbarkeitstheorie ist varianzanalytisch angelegt. Was in der Varianzanalyse ‚Faktor‘ genannt wird, heißt jetzt ‚Facette‘. ‚Bedingungen‘ bezeichnen die ‚Stufen‘ oder ‚Ausprägungen‘ einer Facette.

Ein **Beispiel** von Nußbaum (1987): Gegeben seien drei Facetten: Schüler (1), Items (J) und Beurteiler (K).

Jede Facette wird genau beschrieben. Facette I bestehe aus gehörlosen Schülern. Facette J bestehe aus einem Itemsatz, der die Fähigkeit zur Aussprache bestimmter Lautverbindungen überprüft. Facette K bestehe aus Sonderschullehrern an Gehörlosenschulen.

Zur Schätzung der Varianzanteile je Facette und ihrer Interaktionen werden je Facette Zufallsstichproben von Bedingungen gezogen.

In dem **Beispiel** werde eine Stichprobe von gehörlosen Schülern (Facette 1), von Items (Facette J) und von Gehörlosenlehrern (Facette K) gezogen.

Die gesammelten Daten dienen zur Erstellung zweier Studien: einer Generalisierbarkeits-Studie (G-Studie) und einer Decisions-Studie (D-Studie).

Die **G-Studie** dient dazu, die gesuchten Varianzkomponenten an Stichproben zu ermitteln. (Sie hat einen analogen Zweck wie die Ermittlung von Kennwerten einer Normstichprobe.)

In einer **D-Studie** werden die Varianzkomponenten, die in der G-Studie ermittelt wurden, für neue Untersuchungen ‚dienstbar‘ gemacht. (Sie ähnelt der Anwendung von Normen auf konkrete Untersuchungsfälle.)

In dem **Beispiel** seien die Varianzkomponenten und die Interaktionen der drei Facetten in einer G-Studie ermittelt. Diese Werte können dann dazu dienen, für eine D-Studie, für eine neue Studie über ‚Schüler, Items, Lehrer‘, Varianz-

komponenten vorauszuschätzen und dabei die Bedingungen der Facetten auch zu variieren.

Man kann etwa schätzen, wie zuverlässig - wie reliabel - mit neuen Items (ähnlicher Art) neue Gruppen von Gehörlosen (ähnlicher Art) beurteilt werden können; wie reliabel neue Gruppen von Gehörlosenlehrern neue Schüler beurteilen können.

In diesem Sinne ermöglicht es die Generalisierbarkeitstheorie, schon ermittelte Genauigkeitswerte („Reliabilitätswerte“) zu verwenden zur Vorausschätzung der Genauigkeit von „Anwendungen“. In dieser Eigenart ist sie verwandt mit der Spearman-Brown-Korrektur, sie geht über deren Möglichkeiten aber weit hinaus.

Nach diesem Hinweis auf eine Erweiterung der klassischen Testtheorie zurück zur vierten Modalität, die Meßgenauigkeit eines Tests zu schätzen: zur Ermittlung der *Reliabilität* als *Konsistenz*.

(IV) Konsistenz

Die Schätzung der Reliabilität als Konsistenz bezeichnet das Ausmaß, in dem von denselben Probanden alle Items in gleicher Weise beantwortet werden.

Konsistenz läßt sich von zwei Ansätzen her konzipieren:

- Sie kann verstanden werden als Erweiterung der Halbierungsreliabilität. Ein Test wird zerlegt in so viele Teile, wie er Items hat. Die Korrelation der „Teile“ wird ermittelt (und hochgerechnet).
- Konsistenz läßt sich auch konzipieren von einem varianzanalytischen Paradigma her.

Der varianzanalytische Ansatz sei im einzelnen erläutert:

1. Probanden werden mit Items 1 ... k gemessen. Für den beobachteten Wert gilt: $X = T + E$.

Nach Axiom 1 wird angenommen, daß der **wahre** Wert (T) je Proband über alle Items hinweg konstant bleibt. Wenn nun der **beobachtete** Wert (X) eines Probanden variiert: wie läßt sich diese Variation dann interpretieren? Sie wird interpretiert als Ausfluß von **Fehlern** (E).

Die Variation des beobachteten Wertes je Proband läßt sich verstehen als **Varianz innerhalb** eines Faktors in einem varianzanalytischen Design (s_{in}^2). Diese „Varianz innerhalb“ dient als Schätzung des Meßfehlers.

Demnach wird gleichgesetzt: $s_{in}^2 = s_E^2$.

2. Wenn **zwischen den Probanden** ebenfalls Varianz auftritt, dann werden darin zwei Komponenten gesehen:

- zum einen der Unterschied zwischen den wahren Werten der Probanden,

- zum anderen der Unterschied der Meßfehler zwischen den Probanden. Somit repräsentiert die **Varianz zwischen** den Probanden (s_{zw}^2) sowohl Varianz des wahren Wertes (s_T^2) wie auch die Varianz des Fehlers (s_E^2):
Demnach wird gleichgesetzt: $s_{zw}^2 = s_T^2 + s_E^2$.

Resümee: Die Festlegungen (1) und (2) erlauben eine **Definition der Reliabilität als Konsistenz**. Es gilt:

$$\begin{aligned} s_T^2 + s_E^2 &= s_{zw}^2 \\ s_T^2 &= s_{zw}^2 - s_E^2 \\ s_E^2 &= s_{in}^2 \\ s_T^2 &= s_{zw}^2 - s_{in}^2 \end{aligned}$$

Einsetzen in die Reliabilitätsformel:

$$r_{tt} = \frac{s_T^2}{s_x^2} = \frac{s_T^2}{s_T^2 + s_E^2}$$

$$r_{tt} = \frac{(s_{zw}^2 - s_{in}^2)}{(s_{zw}^2 - s_{in}^2) + s_{in}^2}$$

Auflösen und Kürzen ergibt:

$$r_{tt} = 1 - \frac{s_{in}^2}{s_{zw}^2}$$

Die Ableitung besagt: In einem varianzanalytischen Design wird die Varianz *innerhalb* der Probanden als Indikator des Fehlerwertes interpretiert, die Varianz *zwischen* den Probanden als Indikator der wahren UND der fehlerhaften Varianz.

3. **Ergänzung:** Die ‚Varianz innerhalb‘ enthält eine Komponente, die zurückgeht auf Unterschiede der Itemanforderungen. Items sind nie völlig homogen, somit auch nicht ihre Anforderungen. Insofern ist es legitim, damit zu rechnen, daß auf unterschiedliche Anforderungen unterschiedliche wahre Fähigkeiten der Probanden (T) unterschiedlich reagieren. Das aber besagt: **Ein Teil der ‚Varianz innerhalb‘ ist nicht als Fehler, sondern als zulässige ‚Varianz zwischen den Items‘ zu interpretieren (Lienert & Raatz, 1994, 198).**

Als **Fehlervarianz** ist nur jener Anteil der ‚Varianz innerhalb‘ zu interpretieren, der übrigbleibt, wenn die ‚Varianz zwischen den Items‘ abgezogen ist. Dieser Anteil ergibt die sogenannte **Restvarianz** (s_r^2), sie allein wird als Fehlervarianz gedeutet. Es gilt die Gleichsetzung: $s_r^2 = s_E^2$. Das Ergebnis, in die Reliabilitätsformel übertragen, lautet:

$$r_{tt} = 1 - \frac{s_r^2}{s_{zw}^2}$$

Berechnungsbeispiel

In Kasten 4.3-9 folgt ein Berechnungsbeispiel.

Kasten 4.3-9: Reliabilität als Konsistenz

Pbn	:	Probanden						
P	:	Test-Score je Pb			(Pb 1	:	P ₁	= 11)
I	:	Summen-Score je Item			(Item 1	:	I ₁	= 12)
N	:	Zahl der Pbn			(N			= 5)
k	:	Zahl der Items			(k			= 6)
Berechnungsschritte: Siehe laufenden Text!								
Items								
Pbn	1	2	3	4	5	6	P	
1	1	0	1	2	3	4	11	
2	1	3	4	5	3	2	12	
3	3	2	1	3	2	1	12	
4	4	4	4	4	4	4	24	
5	3	2	1	4	5	3	18	
1	12	11	11	18	17	14	83	

Für die Berechnung gilt:

$$\begin{aligned}
 (\Sigma X)^2 &: \text{Gesamtsumme der Test-Score+ quadriert, hier: } 83^2 = \mathbf{6889} \\
 \Sigma X^2 &: \text{Summe der quadrierten Item-Scores, hier: } 1^2 + \dots + 5^2 + 3^2 = \mathbf{283} \\
 \Sigma P^2 &: \text{Summe der quadrierten Test-Scores, hier: } 11^2 + \dots + 18^2 = \mathbf{1489} \\
 \Sigma I^2 &: \text{Summe der quadrierten Item-Scores, hier: } 12^2 + \dots + 14^2 = \mathbf{1195}
 \end{aligned}$$

Für die folgenden drei Berechnungsschritte A, B C gelten die Kürzel:

$$\begin{aligned}
 QS_{\text{tot}}, s_{\text{tot}}^2 &: \text{Quadratsumme, Varianz } \mathbf{total} \\
 QS_{\text{in}}, s_{\text{in}}^2 &: \text{Quadratsumme, Varianz } \mathbf{innerhalb Pbn} \\
 QS_{\text{zw}}, s_{\text{zw}}^2 &: \text{Quadratsumme, Varianz } \mathbf{zwischen Pbn} \\
 QS_{\text{zl}} &: \text{Quadratsumme } \mathbf{zwischen Items} \\
 QS_{\text{r}}, s_{\text{r}}^2 &: \text{Quadratsumme, Varianz } \mathbf{Rest}
 \end{aligned}$$

(A) Berechnung der **Quadratsummen**:

$$QS_{\text{tot}} = \Sigma X^2 - \frac{(\Sigma X)^2}{N \cdot k} = 283 - \frac{83^2}{5 \cdot 6} = \mathbf{53.4}$$

$$QS_{\text{zw}} = \frac{\Sigma P^2}{k} - \frac{(\Sigma X)^2}{N \cdot k} = \frac{1489}{6} - \frac{83^2}{5 \cdot 6} = \mathbf{18.6}$$

$$QS_{\text{in}} = \Sigma X^2 - \frac{\Sigma P^2}{k} = 283 - \frac{1489}{6} = \mathbf{34.8}$$

$$QS_{zl} = \frac{\Sigma I^2}{N} - \frac{(\Sigma X)^2}{N \cdot k} = \frac{1195}{5} - \frac{83^2}{5 \cdot 6} = 9.4$$

$$QS_r = QS_{in} - QS_{zl} = 34.8 - 9.4 = 25.4$$

(B) Berechnung der **Mittleren Quadratsummen**, also der **Varianzen**:

$$s_{zw}^2 = \frac{QS_{zw}}{N - 1} = \frac{18.6}{6 - 1} = 0.70$$

$$s_{in}^2 = \frac{QS_{in}}{N(K - 1)} = \frac{34.8}{5(-1)} = 1.39$$

$$s_r^2 = \frac{QS_r}{(N - 1)(k - 1)} = \frac{25.4}{(5 - 1)(6 - 1)} = 1.27$$

(C) Berechnung der **Konsistenz**:

Die berechneten Varianzen werden eingesetzt in die Konsistenzformel:

– **„Varianz innerhalb“** (s_{in}^2) als *Fehlerschätzung*:

$$r_{tt} = 1 - \frac{s_{in}^2}{s_{zw}^2} = 1 - \frac{1.39}{4.65} = 0.70$$

– **„Restvarianz“** (s_r^2) als *Fehlerschätzung*:

$$r_{tt} = 1 - \frac{s_r^2}{s_{zw}^2} = 1 - \frac{1.27}{4.65} = 0.73$$

Steht als Fehlervarianz die „Restvarianz“, dann liegt in diesem Falle die Konsistenz geringfügig höher, als wenn die „Varianz innerhalb“ die Fehlervarianz vertritt. In anderen Fällen kann das Gegenteil eintreten.

Probleme der Konsistenzschätzung

Nur wenn die Items homogen sind, läßt sich die Variation innerhalb der Probanden als Fehler interpretieren.

Bedingungsvariation, wie sie bei Retest und Paralleltest erfasst wird, geht in die Messung nicht ein.

4.3.2.5 Test-Score und Vertrauensbereich

Im Rahmen der Reliabilität stellt sich die Frage nach der Präzision eines erreichten Test-Scores. Das Problem sei veranschaulicht an einem Einzelfall:

Ein Proband erreiche im IST 70 einen Gesamtstandardwert (SW) von 107. Der ‚IST 70 Standardwert von 107‘ ist nicht fehlerfrei, wie seine Reliabilität beweist, die zwar hoch liegt, doch unter Eins bleibt. Dann aber lassen sich zwei Fragen stellen:

- Angenommen, *der SW 107 ist der wahre Wert* des Probanden: mit welchen Varianten des Test-Scores ist zu rechnen, wenn die Messung wiederholt wird?
- Oder aber angenommen, *der SW 107 ist nicht der wahre Wert*, sondern eine Summe aus wahren und fehlerhaftem Wert: in welchem Suchbereich liegt dann der wahre Wert?

Es gibt zwei Wege, den Bereich einzugrenzen, in dem der wahre Wert liegt: Bestimmung des

1. Standardmeßfehlers,
2. Standardschätzfehlers.

In beide Bestimmungen geht die Reliabilität ein. Wäre $r_{tt} = 1.00$, dann wäre in beiden Fälle der ermittelte Test-Score gleich dem wahren Wert.

(1) Standardmeßfehler

Bei wiederholten Messungen streuen die empirischen Werte um den wahren Wert. Angenommen etwa, der SW 107 sei der wahre Wert, dann ist damit zu rechnen, daß bei Meßwiederholungen die beobachteten Werte um SW 107 streuen. Diese Streuung ist zu interpretieren als Fehlervarianz. Für diese Fehlervarianz enthält die Reliabilitätsdefinition einen Term, nämlich s_E^2 (S. 73).

$$r_{tt} = 1 - s_E^2 / s_x^2$$

Aufgelöst nach s_E^2 , ergibt sich:

$$s_E^2 = s_x^2 \cdot (1 - r_{tt})$$

Im Dienste einer vereinfachten Schreibweise seien **Kürzel** vereinbart:

- **SE** für die Standardabweichung der Fehlervarianz, also für $\sqrt{s_E^2}$ und
- **SX** für die Standardabweichung der beobachteten Varianz, also für $\sqrt{s_x^2}$.

Dann gilt:

$$SE = SX \sqrt{1 - r_{tt}}$$

Dieser sogenannte **Standardmeßfehler** (SMF) erlaubt es, den Vertrauensbereich zu schätzen, innerhalb dessen bei gegebenem beobachtetem Wert, dem Test-Score, der wahre Wert liegt.

Da angenommen wird, daß der Standardmeßfehler sich normal um den wahren Wert verteilt, dient die Normalverteilung dazu, den **Vertrauensbereich** zu schätzen. Es gilt:

$$VB = CL = X \pm z_{\alpha} \cdot SX \cdot \sqrt{1 - r_{tt}}$$

Es bedeuten:

VB : Vertrauensbereich,

CL : Confidential Limits,

X : beobachteter Wert, Test-Score,

z_{α} : z-Wert für Alphafehler, also gewählte Restwahrscheinlichkeit,
(z.B. $p \leq 5\% \Rightarrow z = 1.96$),

SX : Standardabweichung der empirischen Varianz,

r_{tt} : Reliabilität des angewandten Verfahrens.

Kasten 4.3-10:

Standardmeßfehler/Beispiel

SW: Standardwert, VB: Vertrauensbereich

Gegeben sind:

IST 70 Gesamtwert: SW = 107

Reliabilität (Retest): $r_{tt} = 0.83$,

Standardabweichung: SX = 10

Restwahrscheinlichkeit: $p = 5\% \Rightarrow z_{\alpha} = 1.96$.

Wie groß ist VB?

$$VB = 107 \pm 1.96 \cdot 10 \cdot \sqrt{1 - .83} = 107 \pm 8.1$$

Der Vertrauensbereich, innerhalb dessen mit $p < 5\%$ der wahre Wert des Probanden liegt, beträgt VB = 107 k8.1, das heißt:

Der VB reicht von SW 98.9 bis SW 115.1.

(2) Standardschätzfehler

Lord & Novick (1974, 64, 152) geben einen anderen Weg an, den Bereich zu bestimmen, in dem der wahre Wert liegt. Aus den Axiomen der klassischen Testtheorie leiten sie eine Beziehung zwischen wahren und beobachteten Werten ab, die sich ausdrücken läßt in der Regressionsgleichung:

$$T' = M_X + r_{tt} (X_i - M_X)$$

Neu sind zwei Terme:

T': wahrer Wert (geschätzt mittels Regressionsgleichung),

M_x: Mittelwert der beobachteten Werte jener Gruppe, welcher der Proband zugehört oder entspricht, z.B. der Normgruppe; in dem Beispiel gilt:
 $M_X = 100$.

Der wahre Wert erweist sich somit als ein Ergebnis zweier Schätzungen: des individuellen Test-Scores X_i und des Gruppenmittelwertes M_X , beide gewichtet mit der Reliabilität. (Liegen mehrere Gruppenmittelwerte M_X vor, lassen sich mehrere - auch voneinander abweichende - Schätzungen von T geben. Darin drückt sich die Stichprobenabhängigkeit der Werte, auch der **wahren** Werte, in der klassischen Testtheorie aus.)

Um den Vertrauensbereich zu bestimmen, in dem der wahre Wert liegt, ist zuerst der **Standardschätzfehler (SSF)** zu berechnen:

$$SSF = SX \cdot \sqrt{r_{tt}} \cdot \sqrt{1 - r_{tt}}$$

Der Vertrauensbereich (VB) bei einer Regression heißt **Mutungsintervall (MI)**; es bestimmt sich hier wie folgt:

$$VB = MI = T \pm z_{\alpha} \cdot SSF$$

Kasten 4.3-11:
Standardschätzfehler/Beispiel

SW : Standardwert

T : geschätzter wahrer Wert

VB : Vertrauensbereich = MI: Mutungsintervall

Wie in Kasten 4.3-10 soll gelten:

- IST 70 Gesamtwert: $SW = 107$
- Reliabilität (Retest): $r_{tt} = 0.83$
- Standardabweichung: $SX = 10$
- Restwahrscheinlichkeit: $p = 5\% \Rightarrow z_x = 1.96$

Wie groß ist VB (das Mutungsintervall)?

$$T' = 100 + 0.83 \cdot (-100) = \mathbf{105.81}$$

$$VB = MI = 105.81 \pm 1.96 \cdot 10 \cdot \sqrt{.83} \cdot \sqrt{1 - .83} = 104.81 \pm 6.07$$

Der Vertrauensbereich, hier genannt Mutungsintervall, innerhalb dessen mit $p < 5\%$ der wahre Wert liegt, beträgt $VB = MI = 105.81 \pm 6.07$; das heißt:

Das Mutungsintervall reicht von SW 99.74 bis SW 111.88.

4.3.2.6 Kritische Differenz

Die Berechnung des Standardmeßfehlers kann zu einer zweiten Frage überleiten: Wie weit müssen zwei Test-Scores auseinanderliegen, damit die Differenz (auf einem gewählten Signifikanzniveau) als erheblich gilt? Eine Berechnungsformel (begründet aus der Berechnung des Standardmeßfehlers) lautet:

$$diff_{(X1-X2)} = z_{\alpha} \cdot SX \cdot \sqrt{2 - (r_{11} + r_{22})}$$

Es bedeuten:

$diff_{(X1-X2)}$: Kritische Differenz zwischen Test-Score X1 und Test-Score X2,

r_{11} : Reliabilität des Tests 1, zu dem X1 gehört,

r_{22} : Reliabilität des Tests 2, zu dem X2 gehört.

Der Abstand zwischen Testwert X1 und Testwert X2 ist signifikant (auf einem gewählten Niveau), wenn er größer ist als die Kritische Differenz $diff_{(X1-X2)}$.

Kasten 4.3-12:
Kritische Differenz/Beispiel

Gegeben sind:

IST 70/Untertest AN (Analogien) : SW, = 115

IST 70/Untertest ME (Merken) : SW,, = 102

Reliabilität : r_{11} (AN) = 0.86

r_{22} (ME) = 0.90

Restwahrscheinlichkeit : P = 5 % $\Rightarrow z_{\alpha} = 1.96$

Ist die Differenz zwischen AN und ME signifikant?

Empirische Differenz zwischen AN u. ME:

Kritische Differenz:

$$\text{diff}_{\text{krit, AN-ME}} = 1.96 \cdot 10 \cdot \sqrt{2 - (.86 + .90)} = 9.6 \text{ SW}$$

Die Differenz zwischen AN und ME ist (auf dem 5 %-Niveau) signifikant, denn die Empirische Differenz von 13 SW ist größer als die Kritische Differenz von 9.6 SW.

**Problematik einer Berechnung
von Vertrauensbereich und kritischer Differenz**

Die Bestimmung des Vertrauensbereiches (VB) und der kritischen Differenz (RD) ist im individuellen Falle nur berechtigt, wenn das Individuum zufällig aus der Stichprobe gegriffen ist.

Denn die Berechnung von VB und KD hängt wesentlich ab von der Reliabilität eines Verfahrens. Die Reliabilität ihrerseits ist eine Gruppencharakteristik. ***Problematisch ist die Applizierung einer Gruppencharakteristik auf ein Individuum.*** „Da die Ungenauigkeit der Messung - die Größe der Fehlerstreuung - bei jeder Person verschieden sein kann und auf der Basis der Reliabilität nur ein ‚Durchschnittsmaß‘ dafür zu erhalten ist, ist eine Aussage für eine bestimmte Person . . . nicht ableitbar“ (Wottawa, 1980, 92).

HINWEIS: Gewichtiger als die psychometrische Bedeutung dürfte **der appellative Charakter** sein. Die Beachtung des Standardmeßfehlers oder der kritischen Differenz kann den Anwender daran erinnern, daß er einen Test-Score nie punktuell als ‚wahren Wert‘ interpretieren, ihn vielmehr als Einzel-Element eines weiteren ‚Suchbereiches‘ verstehen sollte.

4.3.3 Validität

Validität betrifft die Frage, wie sich vom Test-Score auf das Ziel-Merkmal, vom numerischen Relativ auf das empirische Relativ schließen läßt. Angesprochen sind mehrere Sachverhalte (Cranach & Frenz, 1969, 305), beispielsweise - die inhaltliche Übereinstimmung einer empirischen Messung mit einem logischen Meßkonzept,

- die Übereinstimmung einer Testmessung mit einer Kriteriumsmessung,
- die Möglichkeit, ein bestimmtes Verhalten vorherzusagen,
- die empirische Verifizierung eines theoretischen Modells,
- vor allem auch ein erkenntnistheoretisches Problem.

Die Fragen, die wir besprechen, lassen sich in drei große Abschnitte gliedern:

- Bestimmung (Definition) von Validität (4.3.3.1),
- Arten von Validität (4.3.3.2),
- Multitrait-Multimethod-Validierung als Paradigma einer Kombination verschiedener Validitätsarten (4.3.3.3).

4.3.3.1 Bestimmung (Definition) von Validität

Bestimmen läßt sich Validität nach „unterschiedlichen diagnostischen bzw. prognostischen Schlußweisen“ (Michel & Conrad, 1982, 55). Zunächst bietet sich eine Zweiteilung an:

1. Vom (beobachtbaren) ‚Verhalten in der Testsituation‘ wird geschlossen auf das (ebenfalls beobachtbare) ‚Verhalten außerhalb der Testsituation‘. Diese Schlußfigur tritt in zwei Varianten auf:

⇒ Ein **Repräsentationsschluß** liegt vor, wenn „das Testverhalten als direkt repräsentativ für ein bestimmtes Gesamtverhalten angesehen“ wird (Michel & Conrad, 1982, 55). Es läßt sich von repräsentativer Validität reden, üblicher ist die Bezeichnung **Inhaltsvalidität**.

⇒ Ein **Korrelationsschluß** liegt vor, wenn ein empirischer Zusammenhang nachgewiesen wird zwischen dem ‚Verhalten in der Testsituation‘ und dem ‚Verhalten außerhalb der Testsituation‘, das seinerseits repräsentiert wird durch ein sogenanntes Kriterium. Daher die Benennung **kriteriumsbezogene Validität!**

Das Paradigma der kriteriumsbezogenen Validität beherrscht die Validierungsprozeduren der klassischen Testtheorie so sehr, daß es auch das **Kürzel für Validität bestimmt hat: r_{tc}** . Das Kürzel zeigt die Korrelation (r) eines Tests (t), mit einem Kriterium (c) an.

Der ‚empirische Zusammenhang‘ wird bei kriteriumsbezogener Validität nach zwei Grundmustern ermittelt:

⇒ als Korrelation mit einem zeitlich koexistentem Kriterium: **Übereinstimmungsvalidität** (concurrent validity) oder

⇒ als Korrelation mit einem (vom Zeitpunkt der Testung her gesehen) zukünftigen Kriterium: **Vorhersagevalidität** oder prädiktive Validität (predictive validity).

2. Vom ‚Verhalten in der Testsituation‘ wird geschlossen auf ‚Fähigkeiten‘, ‚Dispositionen‘ oder ‚Persönlichkeitsmerkmale‘ als Grundlagen oder Bedingungen des Verhaltens. Geschlossen wird somit (nicht auf beobachtbares Verhalten, sondern) auf unbeobachtbare ‚Konstrukte‘. Darum hat sich für diesen Aspekt die Bezeichnung **Konstruktvalidität** eingebürgert.

Die drei Paradigmen einer Validierung - orientiert an Inhalt, Kriterium oder Konstrukt - sind nicht gleichen Ranges. Systematisch ist die inhaltliche Validität vorgeordnet. Warum?

- Für kriteriumsbezogene Validität gilt: Test I wird validiert an Kriterium I. Kriterium I muß seinerseits valide sein, also gemäß diesem Paradigma validiert an einem Kriterium II. Valide muß auch Kriterium II sein, validiert demzufolge an einem Kriterium III. Soll diese Kette nicht zu einem regressum in infinitum führen (also keine Erklärung liefern), setzt wenigstens **ein** ‚Kettenglied‘ ein ‚anderes Validierungsparadigma‘ voraus.
- Für konstruktbezogene Validierung gilt, daß sie Inhalts- und Kriterienvalidität umfaßt. Damit aber schließt sie auch die Kriterienproblematik ein. Sie setzt demnach ebenfalls ein ‚anderes Validierungsparadigma‘ voraus. Als ‚anderes Validierungsparadigma‘ steht nur die Inhaltsvalidierung zur Verfügung. An ‚irgendeiner Stelle‘ in der kriteriums- und konstruktbezogenen Validierungskette ist (unter systematischer Sicht) eine inhaltliche Validierung unumgänglich.

4.3.3.2 Arten von Validität

Die drei Arten von Validität, die eingeführt wurden, seien im einzelnen besprochen:

- Inhaltsvalidität (4.3.3.3.1),
- Kriteriumsbezogene Validität (4.3.3.3.2),
- Konstruktvalidität (4.3.3.3.3).

4.3.3.2.1 Inhaltsvalidität

Inhaltliche Validität ist dann gegeben, wenn der Inhalt der Test-Items das Ziel-Merkmal hinreichend genau definiert. „Aufgrund logischer und fachlicher Überlegungen“ wird die Bedeutung des Ziel-Merkmals bestimmt (Michel & Conrad, 1982, 57; Moser, 1987).

Das zentrale Problem ist die ‚Definition des Ziel-Merkmals. Sie wirft all jene Fragen auf, die unter dem Titel der „Konzeptualisierung von Fragestellung und Testmerkmal“ schon berührt worden sind (S. 32). Das Problem sei hier begrenzt auf den Entwurf charakteristischer Items.

Wünschenswert wäre eine enumerative Lösung: eine vollständige Auflistung relevanter Items. Dieser Weg ist in der Regel versperrt. Das Universum der Items, die zu einem möglichen Test gehören, ist kaum exakt bestimmbar.

Die Praxis muß sich mit einer Näherungslösung begnügen: mit einer exemplarischen Aufzählung von Items, welche die „mengenstiftenden Merkmale“ der Items anführt und an Beispielen verdeutlicht (Michel & Conrad, 1982,

57). Die Näherungslösung setzt voraus, daß es gelingt, Regeln anzugeben, die den Entwurf und die Zusammenstellung von Items leiten.

Oft werden Experten um Mitarbeit an solchen ‚logischen und fachlichen Überlegungen‘ gebeten. Als Experte gilt jemand, der mit dem Ziel-Merkmal vertraut sind.

„Im allgemeinen ist es nicht üblich und häufig auch nicht möglich, Inhaltsvalidität numerisch zu bestimmen. Sie wird vielmehr aufgrund logischer und fachlicher Überlegungen mit oder ohne fachliche Einschränkungen akzeptiert oder verworfen“ (Michel & Conrad, 1982, 57).

Doch gibt es Versuche, die Inhaltsvalidität auch numerisch darzustellen. Differenzierte Verfahren beschreiben Fricke (1974) und Klauer (1984, 1987) im Rahmen kriterienorientierter Leistungsmessung‘ (Kap. 5, S. 129). Es folgt ein Beispiel für eine numerische Darstellung inhaltlicher Validität.

Auch wenn gilt, daß unter systematischer Perspektive inhaltliche Validierung den ersten Rang einnimmt, so ist doch festzuhalten: Ein Testautor sollte sich mit inhaltlicher Validierung nicht zufrieden geben, sondern sich um kriteriumsbezogene Validierung bemühen (Guion, 1977; Pawlik, 1976, 31).

Demonstration:

Beispiel für eine numerische Darstellung der Inhaltsvalidität

Es sei ein einfaches Beispiel für eine numerische Darstellung der Inhaltsvalidität gewählt – in Anlehnung an Dieterich (1973, 45-51; 112).

Aufgabe:

- Ein Merkmal ABC wird möglichst scharf umschrieben.
- Es werden *sechs* Items formuliert, welche das Merkmal ABC in unterschiedlicher Genauigkeit ‚abbilden‘.
- Vier Experten werden mit dem Merkmal ABC und mit den sechs Items vertraut gemacht.
- Die Experten schätzen, wie genau die Items das Merkmal wiedergeben. Die Urteile variieren von 1 bis 5.
 ⇒ Pol 5 bedeutet: ‚sehr zutreffende Wiedergabe‘.
 ⇒ Pol 1 bedeutet: ‚sehr unzutreffende Wiedergabe‘.

Berechnung: Wir gliedern die Berechnung in zwei Schritte:

- Teil 1 bezieht sich auf den *Gesamttest*,
- Teil 2 dagegen auf die sechs *Einzelitems*.

Teil 1: Gesamttest

Die Übereinstimmung der vier Experten läßt sich auf vielfache Weise darstellen, beispielsweise mit dem Konkordanzkoeffizienten von Kendall oder mit dem Ü-Koeffizienten von Fricke (1973, 1974).

Hier sei die Übereinstimmung geprüft in Analogie zur Konsistenzschätzung; das Kürzel r_4 bezeichne die Übereinstimmung zwischen den 4 Experten: Bezogen auf die Items, wird die ‚**Varianz innerhalb**‘ (s_{in}^2) als Indikator der Fehler interpretiert, die ‚**Varianz zwischen**‘ (s_{zwi}^2) als Indikator der fehlerhaften und der wahren Varianz (vgl. S. 86). Demnach gilt die Formel:

$$r_k = 1 - \frac{s_{in}^2}{s_{zwi}^2}$$

Wir bieten zuerst die Daten: in Tabelle A.

Tabelle A

Tabelle A repräsentiert die Urteile der Experten 14 zu den Items a-f,

Es bedeutet: ΣI : Summe der Urteile X je Item über alle Experten.

Items	Experten				ΣI
	1	2	3	4	
a	3	4	5	4	16
b	5	4	3	3	15
c	1	2	1	2	6
d	2	1	2	2	7
e	5	4	5	3	17
f	1	1	2	2	6

Aus den Werten der Tabelle A seien die beiden Varianzen (‚innerhalb‘ und ‚zwischen‘) ermittelt und in die Konsistenzformel eingesetzt. Es ergibt sich:

$$r_4 = 1 - (0.56/7.14) = 0.92$$

Dieses Ergebnis von $r_4 = 0.92$ spricht

- für eine hohe Übereinstimmung der vier Experten (r_4) in ihrer Zuordnung von Inhalten (Items) zu einem Merkmal (ABC) und
- in diesem Sinne für eine hohe Inhalts-Validität des Merkmals ABC (aber auch für eine hohe **Interpretations-Objektivität**).

Teil 2: Einzelitems

Die ‚inhaltliche Merkmalssättigung‘ der einzelnen Items läßt sich ebenfalls auf verschiedene Weise ermitteln, zum Beispiel nach einem Proportionsansatz (Fricke, 1973,50; Mees, 1977, 45) oder nach einem Differenzenansatz (Fisseni & Fermekels, 1995, 135).

Hier sei in Analogie zur Ermittlung des Schwierigkeitsindex verfahren (S. 43): Die **erreichten** Punkte ($\Sigma X = \Sigma I$) werden verglichen mit den **erreichbaren** Punkten ($XX_{\max} = 4 \cdot 5$). Je mehr ‚Punkte‘ die Auswerter attribuieren, für desto „zutreffender“ halten sie das Item (1: sehr unzutreffend, 5: sehr zutreffend). Demnach gilt die Formel:

$$p = \frac{\sum x}{\sum x_{\max}}$$

Einsetzen der Werte aus Tabelle A ergibt für Item a: $p_a = 16/20 = 0.80$. Für alle sechs Items gibt Tabelle B die Werte für p.

Tabelle B						
Items						
	a	b	c	d	e	f
P	.80	.75	.30	.35	.85	.30

Nach Meinung der Experten bilden die Items a, b und e das Merkmal ABC zutreffender ab als die Items c, d und f. In diesem Sinne sind die Items a, b und e inhaltsvalider als die Items c, d und f.

HINWEIS: Das Beispiel dürfte veranschaulichen, daß sich inhaltliche Validität weitgehend der Objektivität annähern kann; denn die Übereinstimmung der Experten läßt sich nicht nur interpretieren als Inhalts-Validität, sondern auch als eine Variante der Interpretations-objektivität,

Zudem nähert sich die ganze Prozedur dem ersten Schritt einer Testkonstruktion: der Konzeptualisierung von Merkmal und Fragestellung (vgl. S. 32). Damit aber ist auch gesagt: Überlegungen zur inhaltlichen Validierung gehören mit zu den ersten Schritten einer Testkonstruktion. In der *kriteriumsorientierten Leistungsmessung* geht die Gesamtkonstruktion (per definitionem) von der inhaltlichen Validierung aus (vgl. Kap 5, S. 129).

4.3.3.2.2 Kriteriumsbezogene Validität

Kriteriumsbezogene Validität wird ermittelt durch Vergleich von Test- und Kriterien-Scores. In der Regel wird der Grad der Übereinstimmung als Korrelationskoeffizient ausgedrückt (r_{tc}). Man unterscheidet

- Übereinstimmungsvalidität (A),
- Vorhersagevalidität (B).

Wir fügen einen Abschnitt an über

- die Grenzen der kriteriumsbezogenen Validität (C).

(A)
Übereinstimmungsvalidität

Übereinstimmungsvalidität besteht in der Übereinstimmung zwischen Test-Scores und solchen Kriteriums-Scores, die gleichzeitig mit den Test-Scores erhoben werden (concurrent validity).

Von denselben Probanden liegen demnach zwei ‚zeitgleiche‘ Meßwertreihen vor: Test-Scores und Kriteriums-Scores, und zwar zeitlich eng beieinander. Der Schluß lautet: ‚Es ist bekannt, welches Verhalten der Kriterium-Score repräsentiert. Darum ist der Test-Score in dem Umfang, den die Höhe des Korrelationskoeffizienten angibt, Indikator für das Merkmal, das im Kriterium zum Ausdruck kommt.‘

Beispiel: Scores, gewonnen mit dem „Grundintelligenztest Skala 3 (CFT 3)“ nach Weiss (und Cattell: 1971), wurden korreliert

- mit Schulnoten und

- mit den Scores anderer Tests, des

⇒ IST (‚Intelligenz-Struktur-Tests‘ von Amthauer 1973),

⇒ PSB (‚Prüfsystems für Schul- und Bildungsberatung‘ von Horn, 1969),

⇒ DWT (‚Differentiellen Wissenstests‘ von Fürntratt, 1969).

Angegeben werden die Höhe des Validitätskoeffizienten (r_{tc}) und die Größe der Stichprobe (N) (vgl. Weise, 1975, 61).

Vergleich	r_{tc}	N
CFT x IST	.68	579
CFT x PSB	.66	61
CFT x DWT	.17-.34	453
CFT x Noten:		
Mathe.: Gymnasium	.53	72
Mathe.: Realschule	.58	90
Mathe.: Berufsschule	.45	78

Ein zentrales Problem stellt sich mit der Frage, auf welche Verhaltensbereiche sich Test- (und Kriteriums-)Aussagen ‚ausdehnen‘ lassen. Eine Faustregel besagt, daß sich eine Generalisierung nur soweit erstrecken darf, wie Test- und Kriteriums-Merkmal einerseits und Ziel-Bereich andererseits einander ähnlich sind. - Diese Frage nach Ähnlichkeit oder Äquivalenz von Situationen wirft dann ihrerseits neue Probleme auf.

„Jeder Test hat so viele Validitätskoeffizienten, wie Kriterien zur Prüfung seiner diagnostischen Leistungsfähigkeit herangezogen werden. Die Höhe und die Unterschiede dieser Koeffizienten hängen nicht nur von der Stärke der Selektionseffekte, sondern auch von der Güte und Eigenart der herangezogenen Kriterien ab“ (Jäger A. O. & Althoff, 1984, 30).

Demonstration:

Numerisches Beispiel für eine kriteriumsbezogene Validierung

Ein vereinfachtes Beispiel soll die numerische Darstellung einer kriteriumsbezogenen Validierung veranschaulichen (vgl. Dieterich, 1973, 112):

- Es sei ein *neuer Test* entwickelt worden, etwa ein Musikalitätstest.

- Es werden *Experten* ausgesucht, denen das Merkmal ‚Musikalität‘ erklärt wird, etwa vier Musiklehrer an einer Realschule.
- Es werden *Probanden* ausgewählt, auf die das Merkmal in abgestufter Ausprägung zutrifft, zum Beispiel die Realschüler, die am Musikunterricht teilnehmen.
- In welchem Maße das Merkmal ‚Musikalität‘ auf die Schüler zutrifft, entscheiden die Musiklehrer (als Experten) *auf einer zehnstufigen Skala*:
 ⇒ *Pol 10 bedeutet: ‚Hohe Merkmalsausprägung‘.*
 ⇒ *Pol 1 bedeutet: ‚Geringe Merkmalsausprägung‘.*

Dieser Kriteriums-Score heiße Y.

HINWEIS: Die Übereinstimmung der vier Lehrer könnte überprüft und als Auswerter-Objektivität interpretiert werden.

- Die Schüler *führen den neuen Musikalitäts-Test durch*. Es können Test-Scores erreicht werden von 1 bis 20.
 ⇒ *Score 20 bedeutet: ‚Hohe Merkmalsausprägung‘.*
 ⇒ *Score 1 bedeutet: ‚Geringe Merkmalsausprägung‘.*

Der Test-Score heiße X

Kasten 4.3-13 gibt ein Zahlenbeispiel.

Kasten 4.3-13:
Kriteriumsbezogene Validität

Numerisches Beispiel: Siehe vorhergehenden Text!

Expertenurteil (Lehrerurteile: Mittelwert)	Y	7	4	8	10	2	6	9
Test-Score	x	11	7	16	17	4	9	13

Ergebnis: Die Korrelation zwischen Kriteriums-Score Y und Test-Score X ergibt: $r_{tc} = 0.95$. Der Koeffizient zeigt eine sehr hohe kriteriumsbezogene Validität an.

HINWEIS: Solchen vereinfachten Prozeduren kommt kein Modellcharakter zu, sie können aber die Möglichkeit veranschaulichen, wie sich Einzelschritte einer kriteriumsbezogenen Validierung kontrollieren lassen.

(B)
Vorhersagevalidität

Vorhersagevalidität besteht in der Übereinstimmung zwischen Test-Scores und solchen Kriteriums-Scores, die später als die Test-Scores erhoben werden (predictive validity).

Bei Übereinstimmungsvalidität wird in der Regel vom Inhalt des Kriteriums auf den Inhalt des Tests geschlossen. Bei Vorhersagevalidität ist das nicht in gleichem Maße möglich: Die Bedeutung des Kriteriumsverhaltens erlaubt kei-

nen unmittelbaren Schluß auf die Bedeutung des Testverhaltens. **Angenommen**, als Prädiktor für Prüfungserfolg erweise sich ein Test, der erfaßt, wie weit ein Proband Prüfungen als Herausforderung interpretiert und nicht als Bedrohung. Die prädiktive Validität werde ermittelt als Korrelation mit der Punktezahl eines Zwischenexamins. Der Inhalt des Zwischenexamins erlaubt nun keinen Schluß auf den Inhalt des Ziel-Merkmals. Festgestellt wird nur, daß sich der getestete ‚Bewältigungsstil‘ als Prädiktor für Hochschulerfolg erwiesen hat.

Beispiel für prädiktive Validität: Test-Scores (von Intelligenz- oder Konzentrationstests), gewonnen vor dem Schulübertritt von Grundschule zur Realschule oder Gymnasium, mögen als Prädiktoren für späteren Schulerfolg dienen (z. B. Mittlere Reife, Abitur).

Das zentrale Problem kriteriumsbezogener Validierung ist die Qualifikation des Kriteriums. „Verständlicherweise muß das gemessene Kriterium selbst ausreichende Reliabilität und Validität aufweisen“ (Michel & Conrad, 1982, 61).

(C) Grenzen kriteriumsbezogener Validität

Die Axiome der klassischen Testtheorie begrenzen die kriteriumsbezogene Validität. Das sei unter drei Titeln erläutert:

- Reliabilitätsindex (I),
- Einfache und Doppelte Minderungskorrektur (II),
- Verdünnungsparadox (attenuation paradox) (III).

(I) Reliabilitätsindex

Der Reliabilitätsindex ist ein Kennwert, der die obere Grenze der kriteriumsbezogenen Validität eines Tests festlegt. Er ergibt sich aus folgender Argumentation: Der engste Zusammenhang, den ein Test eingehen kann, ist der zu seinen wahren Werten. Also sei die Korrelation zwischen beobachteten und wahren Werten ermittelt: $r_{t,T}$.

$$(I) \quad r_{t,T} = \frac{cov_{t,T}}{s_t \cdot s_T}$$

Es bedeuten:

t : beobachteter Testwert,
 s_t^2, s_t : Varianz, Standardabweichung der beobachteten Testwerte,
 s_T^2, s_T : Varianz, Standardabweichung der wahren Werte,
 $cov_{t,T}$: Kovarianz der beobachteten und der wahren Werte.

Aus Axiom 3 hatte sich ergeben, daß gilt: $\text{cov}_{t,T} = s_T^2$ (S. 72). Einsetzen im Zähler von (1) erbringt:

$$r_{t,T} = \frac{s_T^2}{s_t \cdot s_T} = \frac{s_T}{s_t}$$

Aus der Ableitung folgt:

- Die Korrelation zwischen beobachteten und wahren Werten bestimmt sich als ein Quotient, der besagt: $r_{t,T} = s_T/s_t$.
- Dieser Term läßt sich verstehen als die Wurzel der Reliabilität, die definiert ist als: $r_{tt} = s_T^2/s_t^2$.

Somit ergibt sich: Die Korrelation zwischen beobachteten und wahren Werten ($r_{t,T}$) entspricht der Wurzel der Reliabilität:

$$r_{t,T} = \frac{s_T}{s_t} = \sqrt{\frac{s_T^2}{s_t^2}} = \sqrt{r_{tt}}$$

Dieser Ausdruck $\sqrt{r_{tt}}$, der **Reliabilitätsindex**, gibt die Obergrenze der kriteriumsbezogenen Validität an, er besagt: Enger als mit den wahren Werten können die beobachteten Werte eines Tests nicht korrelieren.

Beispiel: Wenn die Reliabilität bei $r_{tt} = 0.64$ liegt, kann die kriteriumsbezogene Validität höchstens den Wert $r_{tc} = 0.80$ erreichen.

Wenn sich die Obergrenze der kriteriumsbezogenen Validität bestimmt als Wurzel des Reliabilitätskoeffizienten, dann folgt daraus: Bei einem Koeffizienten, der höher liegt, handelt es sich um ein Artefakt.

(II) Einfache Minderungskorrektur

Wenn die kriteriumsbezogene Validität ihre Obergrenze nicht erreicht, läßt sich fragen: Worauf geht die Absenkung zurück? Eine Antwort ergibt sich, indem geklärt wird, wie hoch die Korrelation ausfällt zwischen **beobachteten** Test- und **wahren** Kriteriumswerten: $r_{t,Tc}$.

$$r_{t,Tc} = \frac{\text{cov}_{t,Tc}}{s_t \cdot s_{Tc}}$$

Es bedeuten:

T_C : wahrer Kriteriumswert,

s_t, s_{Tc} : Standardabweichung der beobachteten Testwerte,
der wahren Kriteriumswerte,

$\text{cov}_{t,Tc}$: Kovarianz von beobachteten Test- und wahren Kriteriumswerten.

Da der wahre Kriteriumswert (T) so wenig beobachtbar ist wie der wahre Testwert (T), ist auch ihre Korrelation ($r_{t,Tc}$) nicht unmittelbar bestimmbar. Doch lassen sich ihre Terme $\text{cov}_{t,Tc}$ und s_{Tc} gemäß den Axiomen der klassischen Testtheorien durch Äquivalente ersetzen, die zu der Formel führen:

$$r_{t,c} = \frac{r_{tc}}{\sqrt{r_{cc}}}$$

Neu ist nur ein Term:

r_{cc} : Reliabilität (der beobachteten Werte des Kriteriums).

Das Ergebnis besagt: Die Korrelation zwischen beobachteten Test-Scores und wahren Kriterien-Scores ($r_{t,Tc}$) läßt sich ermitteln als Quotient, der zwei empirisch bestimmbare Größen enthält:

- im Zähler die ‚beobachtete‘ Validität: r_{tc} ,
- im Nenner die Wurzel der Kriteriumsreliabilität: $\sqrt{r_{cc}}$.

Diese Formel ermöglicht eine Hochrechnung ‚a posteriori‘: die sogenannte **einfache Minderungskorrektur**, die von der ‚Fiktion‘ eines vollständig reliablen Kriteriums ausgeht.

Die Kriteriumsreliabilität r_{cc} erhält die Rolle eines Gewichtes. Denn es gilt:

- *Je mehr die Kriteriumsreliabilität r_{cc} ansteigt, desto mehr nähert sich $r_{t,Tc}$ der ‚beobachteten Validität‘ r_{tc}*
- *Je weiter jedoch die Kriteriumsreliabilität r_{cc} abfällt, desto mehr steigt $r_{t,Tc}$ an.*

Dies besagt: Einer der Gründe dafür, daß die Validität eines Tests ihre Obergrenze nicht erreicht, liegt in der geminderten Reliabilität des Kriteriums (r_{cc}). Theoretische Ableitungen gemäß den Axiomen der klassischen Testtheorie ermöglichen es, den ‚Verlust an Validität‘ zu schätzen, den eine Minderung der Kriteriumsreliabilität ‚verursacht‘. Diese Schätzung führt zu einer Hochrechnung, die (von der Fiktion eines absolut reliablen Kriteriums ausgeht und so) den Verlust ausgleicht. - Den Effekt der Hochrechnung soll ein Beispiel in Kasten 4.3-14 veranschaulichen, aber erst, nachdem auch die doppelte Minderungskorrektur besprochen worden ist.

(III) Doppelte Minderungskorrektur

Ein weiterer Grund für eine geringe Validität kann die Unreliabilität des **Tests** sein. Um auch diesen Mangel zu berücksichtigen, berechnet man die Korrelation zwischen wahren Test- und wahren Kriteriumswerten: $r_{T,Tc}$. Gemäß den Axiomen der klassischen Testtheorie ergibt sich die Formel:

$$r_{t,Tc} = \frac{r_{tc}}{\sqrt{r_{cc} \cdot r_{tt}}}$$

Ein **Beispiel** in Kasten 4.3-14 soll den Effekt der einfachen und doppelten Minderungskorrektur zeigen.

Kasten 4.3-14:
Einfache und doppelte Minderungskorrektur

Gegeben sind <i>drei Tests: A, B und C.</i>				
Bekannt sind		Ermittelt werden die Koeffizienten		
- die ‚beobachtete‘ Validität der Tests A, B, C	r_{tc}	- nach einfacher Minderungskorrektur	: $r_{t,Tc}$	
- die Reliabilität des Kriteriums	r_{cc}	- nach doppelter Minderungskorrektur	: $r_{T,Tc}$	
- die Reliabilität der Tests A, B, C	r_{tt}			
	Tests	A	B	C
r_{tc} Validität des Tests		.48	.34	.37
r_{cc} Reliabilität des Kriteriums		.55	.80	.60
r_{tt} Reliabilität des Tests		.70	.85	.89
$r_{t,Tc}$		$\frac{.48}{\sqrt{.55}}$	$\frac{.34}{\sqrt{.80}}$	$\frac{.37}{\sqrt{.60}}$
Einfache Minderungskorrektur		.64	.39	.46
$r_{T,Tc}$		$\frac{.48}{\sqrt{.55} \cdot .70}$	$\frac{.34}{\sqrt{.80} \cdot .85}$	$\frac{.37}{\sqrt{.60} \cdot .89}$
Doppelte Minderungskorrektur		.77	.41	.51

Erläuterung zu Kasten 4.3-14: Die Minderungskorrekturen werten die Validitätskoeffizienten auf. Der Effekt ist um so größer, je geringer die Reliabilität des Kriteriums (r_{cc}) und die des Tests (r_{tt}) ist.

Für die Praxis der Testpsychologie sind solche korrigierten Validitätskoeffizienten wenig dienlich, weil sie keine empirischen Zusammenhänge aufhellen, vielmehr theoretische Systemstrukturen abbilden und von ‚Fiktionen‘ ausgehen (Michel & Conrad, 1982, 61). In erster Linie tragen sie zum Verständnis der klassischen Testtheorie bei. Diese Rolle soll der folgende Abschnitt verdeutlichen.

Partielle Inkompatibilität von Reliabilität und Validität

Vom Konzept der klassischen Testtheorie her sollte man erwarten: Höhere Reliabilitäten (von Test und Kriterium) erbringen auch höhere Validitäten. (So legt es beispielsweise der Reliabilitätsindex nahe: Je höher der Reliabilitätskoeffizient selber, desto höher auch der Reliabilitätsindex, also die Obergrenze der Validität.) Die beiden Minderungskorrekturen **können** ein Spannungsverhältnis zwischen Reliabilität und Validität anzeigen.

Denn hält man in der Formel der Doppelten Minderungskorrektur den Zähler **konstant**, also die ‚beobachtete‘ Validität (r_{tc}) und **erhöht** den Nenner, also die Reliabilität von Test und Kriterium (r_{tt} , r_{cc}), dann **sinkt die hochgerech-**

nete Validität ($r_{T,Tc}$) ab. Ein Beispiel in Kasten 4.3-15 soll den Effekt verdeutlichen.

Kasten 4.3-15:
Effekt der Minderungskorrektur: Verdünnungsparadox

Um den Effekt der doppelten Minderungskorrektur zu veranschaulichen, soll gelten:				
- Die ‚beobachtete‘ Validität bleibt von Test A zu Test C gleich:				(r_{Tc}).
- Die Kriterien-Reliabilität steigt von A zu C hin an:				(r_{cc}).
- Ebenso steigt die Test-Reliabilität von A zu C:				(r_{Tt}).
Effekt: Steigt die Reliabilität des Kriteriums (r_{cc}) und steigt die Reliabilität des Tests (r_{Tt}), dann sinkt die doppelt korrigierte Validität ($r_{T,Tc}$).				
Tests \Rightarrow		A	B	c
Validität des Tests:	r_{Tc}	.50	.50	.50
Reliabilität des Kriteriums:	r_{cc}	.60	.70	.80
Reliabilität des Tests:	r_{Tt}	.60	.70	.80
Doppelte Minderungskorrektur: $r_{T,Tc}$.83	.71	.62

Erläuterung zu Kasten 4.3-15: Es enthüllt sich eine paradoxe Konsequenz, genannt **Verdünnungsparadox** (attenuation paradox):

- **Steigt** die Reliabilität von Kriterium und Test ($A \Rightarrow C$), dann **sinkt** die Korrelation zwischen wahren Test- und wahren Kriteriumswerten ($r_{T,Tc}$).
- **Sinkt** die Reliabilität von Kriterium und Test ($C \Rightarrow A$), dann **steigt** die Korrelation zwischen wahren Test- und wahren Kriteriumswerten ($r_{T,Tc}$).

Konsequenz: Hohe Reliabilität und hohe Validität sind gemäß der Axiomatik der klassischen Testtheorie partiell unvereinbar. Zwar ist kriteriumsbezogene Validität die dominante Form von Validität im Rahmen der klassischen Testtheorie. Gerade sie eignet sich aber auch dazu, Grenzen des ‚klassischen‘ Ansatzes aufzuzeigen.

Noch einmal zur Beziehung von Reliabilität und Validität: „Aus der Beziehung zwischen Reliabilität und Validität ergibt sich im übrigen, daß die in der Literatur fast durchweg gestellten Anforderungen an die Reliabilität von Tests überspitzt sind. Die . . . immer wieder gestellte Forderung, daß sich Reliabilitätskoeffizienten um oder über .90 bewegen sollten, steht in einem krassen Mißverhältnis zu den praktisch erreichten Validitätskoeffizienten, die nur selten über .60 liegen. Es muß deshalb mit Nachdruck wiederholt werden, was Guilford bereits 1946 ausführte: ‚Relativ zu viel Aufmerksamkeit wird der Reliabilität und zu wenig der Validität geschenkt . . . Eine hohe Reliabilität sollte nie als selbständiges Ziel erstrebt werden, Sie ist nur insoweit wichtig, als sie zur Validität beiträgt‘ “ (Michel & Conrad, 1982, 53-54).

4.3.3.2.3 Konstruktvalidität

Konstruktvalidität heißt die dritte Validitätsklasse. Umschreiben läßt sie sich als Übereinstimmung zwischen Test-Score und einem Netz anderer Scores oder

anderer Aussagen. Worin besteht die Besonderheit? Das Konstrukt, dessen Indikator das Testverhalten ist, wird eingebettet in ein sogenanntes nomologisches Netz' theoretisch verwandter oder theoretisch entfernter Konstrukte.

Für die konkrete Durchführung der Validierung liefert die Konzeption keine Handlungsanweisung. Sie formuliert ein Prinzip, ein Programm, aber keinen Imperativ für Einzelschritte. Darum sollte sie auch eher im Sinne eines Prozesses verstanden werden, weniger in dem eines Zustandes, also eher als Validierung denn als Validität. Inhaltliche und kriteriumsbezogene Validität schließt sie ein.

Ein klassisches Instrument der Konstruktvalidierung ist die **Faktorenanalyse**. Als heuristische Prozedur kann sie in gegebenen numerischen Relativen (ihrerseits Abbildern empirischer Relative) ‚Strukturen‘ erkennen helfen: Sie kann Items identifizieren, die auf demselben Faktor hoch laden. Solche ‚gleichartigen‘ Items lassen sich (möglicherweise) als Indikatoren einer gemeinsamen Eigenschaftsdimension interpretieren und - gemäß der internalen Konstruktionsstrategie - zu einer Skala zusammenfassen. Dazu Fischer (1974, 77):

„Da die Faktorenanalyse eine der gängigsten Methoden zur Auffindung von latenten Eigenschaftsdimensionen ist, und da es sich bei ihr zudem um eine Verallgemeinerung der klassischen Testtheorie handelt,“ läßt sich „das faktorenanalytische Modell zur Präzisierung des Begriffes der Konstruktvalidität“ heranziehen.

Es ist kein Vorgehen angebbbar, das allein Konstruktvalidität verbürgte. Ein solcher Rang kommt auch der Faktorenanalyse nicht zu. Doch dürften sich **drei Aspekte** ausgliedern lassen, die jede Art von Konstruktvalidierung einschließt (vgl. Dieterich, 1973, 135-136; Lienert, 1969, 262-263):

- Konstrukte und ihr Theoriebezug (A),
- Analyse der Konstrukt-Items (B),
- Zusammenhang zwischen Testkonstrukt und anderen Konstrukten (C).

(A)

**Das Testkonstrukt wird aus einer Theorie abgeleitet
oder einer Theorie zugeordnet.**

Beispiele: „Resignation und Dominanz“ lassen sich aus der Theorie Adlers ableiten und in Items abbilden. Ein Fragebogen, der Resignation und Dominanz erfassen soll, kann auf diese Weise theoriegeleitet konstruiert werden.

Es lassen sich auch Items sammeln, die (nach Augenschein) das Konstrukt Aggression betreffen. Die genauere Fassung und Formulierung der Items und der Skalen könnte sich dann an einer Aggressionstheorie orientieren, etwa von Bandura (1973) oder von Berkowitz (1962) (vgl. S. 187).

Ableitung und Einpassung dürften sich bei einer Konstruktvalidierung ergänzen. So leitete Jackson, D. N. (1974) die Konzepte des Fragebogens ‚Personality Research Form‘ (PRF) aus dem System von Murray ab, modifizierte und präziserte sie dabei jedoch erheblich.

(B)
**Die Items, die das Konstrukt repräsentieren,
 werden sowohl semantisch als auch statistisch analysiert.**

Beispiele solcher Analysen könnten sein:

- Inhaltliche Analysen (Welchen Aspekt einer Theorie repräsentiert ein Item? Welche Items bilden ähnliche Inhalte ab? usw.),
- sprachliche Verdeutlichungen (Sind die Items stichprobenbezogen verständlich? Vermeiden sie doppelte Verneinungen? Sind sie für Experten eindeutig? Bringen parallel formulierte Items neue Aspekte in den Test? usw.),
- Iteminterkorrelation,
- Korrelation der Items mit ‚fremden‘ Skalen,
- Itemanalysen,
- Faktoren-, Clusteranalysen der Items
- usw.

(C)
**Der Zusammenhang des Testkonstrukts mit anderen Konstrukten
 wird untersucht.**

Der Zusammenhang des Testkonstrukts mit anderen Konstrukten wird untersucht, und zwar

- sowohl zu Variablen, zu denen *Konvergenz*,
- als auch zu Variablen, zu denen *Diskordanz* angenommen wird.

Beispiele für solche Untersuchungen könnten sein:

- Interkorrelation von Subskalen desselben Tests,
- Korrelation der ‚neuen‘ Skala mit ähnlichen Skalen,
- Korrelation der ‚neuen‘ Skala mit ‚unähnlichen‘ Skalen,
- Korrelation mit Kriterien, die das Ziel-Merkmal repräsentieren,
- Korrelation mit Kriterien, die andere Merkmale repräsentieren
- usw.

Beispiel: Bei der Validierung des „IPC-Fragebogens zu Kontrollüberzeugungen“ ermittelte Krampen (1981, 12-16) die Beziehung seiner Skalen zu theoretisch verwandten und zu theoretisch entfernten Konstrukten. (‚Kontrollüberzeugung‘ betrifft die Frage, ob ein Individuum sein Verhalten stärker bestimmt sieht von Faktoren innerhalb oder außerhalb der eigenen Person.)

Die Buchstaben I, P, C bezeichnen drei Skalen:

- I: Gefühle, selber sein Verhalten zu kontrollieren (**Internal** control'),
 P: Gefühle der Abhängigkeit von mächtigen anderen (**Powerful** others'),
 C: Gefühle der Abhängigkeit vom Schicksal (**Chance**').

Beispiele seien nur für Skala C (Chance) genannt:

- Im Sinne der **Konvergenz** werden für C (Chance: 'Abhängigkeit vom Schicksal') höhere Korrelationen berichtet zu verwandten Konstrukten wie Hoffnungslosigkeit, Depressivität, Rigidität usw.
- Im Sinne der **Diskordanz** werden niedrigere Korrelationen berichtet zu entfernten Konstrukten wie Aggressivität, Geselligkeit, Gelassenheit usw.

Resümee zur Konstruktvalidierung: *Die Konstruktvalidierung bringt vor allem einen Gewinn. Sie bettet das pragmatische Vorgehen der Kriteriumsvalidierung in wissenschafts- und meßtheoretische Überlegungen ein, sie leitet an, persönlichkeits-theoretische Konzepte und praktisch-psychologische Maßnahmen zu verknüpfen. „In dieser umfassenden Sichtweise schließt Konstruktvalidität als Oberbegriff alle anderen Validitätsarten ein“ (Michel & Conrad, 1982, 71).*

In der Möglichkeit, das Vorgehen in großem Maße beliebig zu gestalten, sehen Kritiker eine Gefahr (Pervin, 1981, 147):

- es fehle der Konstruktvalidierung an methodischer Strenge, sie leite nicht an zu einer präzisen Fassung der Konstrukte,
- sie verführe zum Vergleich von Konstrukten, deren Bedeutungsäquivalenz unzureichend geklärt sei.

Der Gewinn dürfte überwiegen, vor allem, wenn man abwägt, welchen Dienst die Kombination verschiedener Validitätsklassen leistet.

4.3.3.3 Multitrait-Multimethod-Validierung:

Paradigma einer Kombination von Validierungsarten

Ein Beispiel für die Möglichkeit, verschiedene Validierungsklassen zu verknüpfen, bietet die sogenannte Multitrait-Multimethod-Validierung von Campbell und Fiske (1959; vgl. Bagozzi, 1993; Byrne & Goffin, 1993; Hubert & Baker, 1978; Ostendorf, Angleitner & Ruch, 1986; Schmitt, Coyle & Saari, 1977; Schmitt & Stults, 1986).

Es sei nur der Grundgedanke entwickelt

Der Ansatz geht von einer Merkmals-Methoden-Einheit (trait-method-unit) aus: Ein Test erfaßt ein Merkmal mit einer bestimmten Methode; er erfaßt kein ‚Merkmal an sich‘. Insofern enthalten die mittels Tests gewonnenen Daten zwei Komponenten: Angaben über Methoden (hier die Methode ‚Test‘) und Angaben über Merkmale (z.B. Dominanz, Depression, Offenheit). Ziel der

Multitrait-Multimethod-Validierung ist es, die beiden Komponenten zu trennen.

Dazu braucht man wenigstens zwei Merkmale und zwei Methoden. Am effektivsten ist eine Trennung, wenn sich die Merkmale deutlich voneinander abheben und wenn auch die Methoden erheblich divergieren.

Um ein **Beispiel** zu entwickeln:

- Gegeben seien drei **Konstrukte** A, B, C, etwa Dominanz, Depression, Offenheit.
- **Gemessen** sei jedes Merkmal mit drei **Methoden** I, II und III (Test, Exploration, Verhaltensbeobachtung).

Konstrukte und Methoden seien einander so zugeordnet, daß jede der drei Methoden (I, II, III) jedes der drei Merkmale (A, B, C) erfaßt.

Bei diesem Plan lassen sich unterschiedliche Arten von Validität konzipieren:

- **Konvergente Validität:** Die Scores für dasselbe Konstrukt, etwa A, sollen über alle drei Methoden hinweg hoch miteinander korrelieren. Das besagt: Gleiche Konstrukte sollten eng miteinander kovariieren, unabhängig von der Methode, mit der sie erfaßt werden.
- **Diskriminante Validität:** Die Scores für verschiedene Konstrukte, etwa für A und B, sollten bei derselben Methode, etwa bei Methode II, niedrig miteinander korrelieren. Das besagt: Gleiche Methode sollte zu geringer Kovariation zwischen unterschiedlichen Konstrukten führen.

Die Zuordnung von Merkmal und Methode wird in einer eigenen Matrix dargestellt (Kasten 4.3-16).

In der Matrix interessieren vor allem die **Diagonalen**:

- Die **Hauptdiagonale** enthält Korrelationen der Merkmale A, B, C mit den Merkmalen A, B, C innerhalb derselben Methode. Es handelt sich demnach um **Reliabilitätskoeffizienten**. Beispielsweise lauten die Koeffizienten von oben links nach unten rechts: 0.79, 0.80, 0.77 . . . 0.83, 0.81, 0.85.
- Die drei **Nebendiagonalen** enthalten Korrelationen der Merkmale A, B, C mit den Merkmalen A, B, C bei verschiedenen Methoden. Es handelt sich demnach um **Validitätskoeffizienten**. Es geht um konvergente Validierung. Beispielsweise lautet

⇒ die Nebendiagonale bei I x II:	0.48, 0.46, 0.51,
⇒ die Nebendiagonale bei II x III:	0.55, 0.65, 0.59.

HINWEIS:

- In den Dreiecken *unterhalb der Hauptdiagonalen* stehen Korrelationen des *einen* Merkmals mit einem *anderen* Merkmal *innerhalb* derselben Methode, beispielsweise A x B bei Methode I (Test): $r = 0.14$.

- In den Dreiecken **unterhalb der Nebendiagonalen** stehen Korrelationen des *eines* Merkmals mit einem *anderen* Merkmal bei *verschiedenen* Methoden: beispielsweise A (Test) x B (Exploration): $r = 0.12$.
- In beiden Fällen geht es um **diskriminante Validierung**.

Weiter sei die Matrix nicht aufgeschlüsselt.

Kasten 4.3-16:
Multitrait-Multimethod-Matrix

Drei Hinweise:										
(1) I, II, III bezeichnen drei Methoden : <i>Test, Exploration, Verhaltensbeobachtung</i> .										
(2) A, B, C bezeichnen drei Merkmale : <i>Dominanz, Depression, Offenheit</i> .										
(3) Jedes Merkmal wird mit jeder Methode erfaßt.										
		I TEST			II EXPLOR.			III VERHALTENSB.		
		A	B	C	A	B	C	A	B	C
I	A	79								
	B	14	80							
	C	06	34	77						
II	A	48			73					
	B	12	46		01	77				
	C	07	19	51	09	12	81			
III	A	48			55			83		
	B	09	49		12	65		10	81	
	C	11	11	56	04	28	59	03	11	85
Koeffizienten der Kürze halber ohne Dezimalpunkt!										

Es seien vier **Forderungen** an Design, Konstrukte und Methoden referiert, die - sind sie erfüllt - eine effektive Multitrait-Multimethod-Validierung versprechen:

- Die **konvergenten Validitäten** (in den Nebendiagonalen) sollen sich signifikant von Null unterscheiden und sollen eine **substantielle Höhe** erreichen.
- Die konvergenten Validitäten sollen **höher** liegen **als die diskriminanten Validitäten bei gleicher Methode** (in den Dreiecken unterhalb der Hauptdiagonalen).
- Die konvergenten Validitäten sollen **höher** liegen **als die diskriminanten Validitäten bei verschiedenen Methoden** (in den Dreiecken unterhalb der Nebendiagonalen).
- Die **Muster der divergenten Validitäten** sollen **äquivalent** sein. (Unabhängig von der Methode sollen die gleichen Muster die gleiche Zuordnung der Konstrukte zueinander spiegeln.)

Die vier Forderungen erweisen sich als plausibel. Sie zeigen aber zugleich auch das **Hauptproblem** der Multitrait-Multimethod-Validierung an: An welchen Kriterien läßt sich ablesen, ob eine Korrelation ‚substantiell‘ ist? Wann liegen die konvergenten Validitäten hoch genug, um sich von den diskrimi-

nanten Validitäten angemessen abzuheben? Bei welchen numerischen Werten gelten die Muster der divergenten Validitäten als äquivalent?

Spätere Autoren haben diese Nachteile aufzuarbeiten versucht. Es hat sich geradezu eine selbständige Forschungsrichtung entwickelt, die diesen wichtigen Validierungsansatz erweitern und vertiefen soll (vgl. Bagozzi, 1993; Byrne & Goffin, 1993; Hubert & Baker, 1978; Ostendorf, Angleitner & Ruch, 1986; Schmitt, Coyle & Saari, 1977; Schmitt & Stults, 1986).

4.4 Normierung oder Eichung

Ein Test, der objektiv, reliabel und valide ist, kann sich als nützliches diagnostisches Instrument bewähren - vor allem, wenn er normiert oder geeicht wird. Eichen oder Normieren nennt man das „Berechnen einer Kennzahl, die das Verhältnis des einzelnen Testwertes zu den Ergebnissen einer Stichprobe zum Ausdruck bringt“ (Wottawa, 1980, 102).

Normen liefern demnach Vergleichswerte. Sie ermöglichen es, anzugeben, welche Position ein Proband einnimmt bezüglich der Werte anderer Personen oder bezüglich eigener Werte.

Zwar liegen Normwerte bereits vor: von den Stichproben, an denen Item- und Testkennwerte ermittelt wurden. Doch empfiehlt es sich neue Kennzahlen zu berechnen - mit dem Ziel, den ‚neuen‘ Test zusätzlich zu überprüfen.

Zwei Fragen seien besprochen: die Berechnung von Normwerten (4.4.1), die *Probleme* einer Normierung (4.4.2).

4.4.1 Berechnung von Normen

Normen als Vergleichswerte lassen sich in verschiedenen Skalen angeben. Zwei dürften die wichtigsten sein:

- Rohwerte (4.4.1.1),
- transformierte Werte (4.4.1.2).

Es folgt ein Hinweis auf mögliche Klassifikationen von Testwerten und auf übliche Normskalen (4.4.1.3).

4.4.1.1 Rohwerte

Rohwerte bestehen in der Zahl der Punkte, die für ‚Lösungen‘ oder für ‚Antworten in Schlüsselrichtung‘ gegeben werden. Sie ermöglichen die Berechnung zentraler Werte, etwa des Mittelwertes und der Standardabweichung. Kasten 4.4-1 gibt ein Beispiel.

Kasten 4.4-1:
Rohwerte als Normen
Eysenck Personality Inventory (EPI, Form A, Eggert, 1983, 26)
M: Mittelwert, SD Standardabweichung, N: Probandenzahl

<i>Stichprobe</i>		<i>Alter</i>	<i>Neurotizismus RW</i>	<i>Extraversion RW</i>	<i>Lügenskala RW</i>
<i>Jugendliche</i>	<i>M</i>	14.1	6.8	12.9	3.9
<i>Männlich</i>	<i>SD</i>	0.6	3.7	3.2	1.9
<i>14 Jahre</i>	<i>N</i>	586.0			
<i>Jugendliche</i>	<i>M</i>	14.0	8.6	13.1	3.8
<i>Weiblich</i>	<i>SD</i>	0.5	4.3	3.6	1.8
<i>14 Jahre</i>	<i>N</i>	524.0			

Die Rohwerte aus Kasten 4.4-1 genügen, um Probanden zu klassifizieren.

Beispiel: Als durchschnittlich seien Testwerte klassifiziert, die in das Intervall ‚Mittelwert plus/minus eine Standardabweichung‘ fallen, als überdurchschnittlich Werte, die oberhalb, und als unterdurchschnittlich Werte, die unterhalb des Durchschnittsintervalls liegen.

Eine 14jährige Probandin erhalte in Extraversion 18 Rohpunkte. Dieser Wert spricht für eine überdurchschnittliche Ausprägung von Extraversion; denn der obere Durchschnittsbereich der Vergleichsgruppe 14jähriger Mädchen geht bis zu $(13.1 + 3.6 =) 16.7$ Rohpunkten. Der Rohwert der Probandin übersteigt diese Grenze.

Ein Nachteil von Rohpunkten liegt darin, daß sich die Werte in verschiedenen Tests unterschiedlich verteilen können, somit auch unterschiedliche Zentralwerte aufweisen und dadurch Vergleiche erschweren. Kasten 4.4-1 veranschaulicht diesen Nachteil:

- Für die Jungen liegt der Mittelwert in *Neurotizismus* bei 6.8, für die Mädchen bei 8.6 Rohpunkten.
- Für die Jungen liegt der Mittelwert in *Extraversion* bei 12.9, für die Mädchen bei 13.1 Rohpunkten.

Diesen Nachteil versucht man auszugleichen, indem man die Rohwerte in Standardwerte transformiert.

4.4.1.2 Transformierte Werte

Rohwerte verschiedener Tests lassen sich auf gleiche Normskalen transformieren. Dabei unterscheidet man drei Hauptklassen:

- Äquivalentnormen (A),
- Variabilitätsnormen (B),
- Prozentrange (C).

(A) Äquivalentnormen

Äquivalentnormen beruhen auf einer Transformation, bei der einem Rohwert ein Zeit-Äquivalent zugeordnet wird, das angibt, welcher Alterstufe eine Testleistung angemessen ist (Dieterich, 1973, 190; Lienert & Raatz, 1994, 282).

Verglichen werden zwei Altersangaben: das Lebensalter und das ‚Leistungsalter‘ -bei Intelligenztest heißt es ‚Intelligenzalter‘ (IA), bei Entwicklungstests ‚Entwicklungsalter‘ (EA).

Beispiel: Ein Intelligenztest gebe für jede Jahreskohorte zwölf Aufgaben vor. Jede Aufgabe repräsentiert demnach einen ‚Intelligenzmonat‘.

Ein Proband, der sieben Jahre und drei Monate alt ist, löse

- alle Aufgaben bis zur Altersgruppe der Sechsjährigen,
- 8 Aufgaben der Sieben-,
- 5 Aufgaben der Acht- und
- 3 Aufgaben der Neunjährigen.

Das ‚Intelligenzalter‘ (IA) beträgt demnach: 6 Jahre + 8 Monate + 5 Monate + 3 Monate = 88 Monate. Das Lebensalter (LA) beträgt 87 Monate.

$$IQ = \frac{IA}{LA} \quad \text{Einsetzen: } IQ = \frac{88}{87} = 1.01$$

Intelligenzalter (IA) und Lebensalter (LA) werden verglichen. Ihr Quotient ergibt den sogenannten ‚Intelligenzquotienten‘ (IQ):

In der Regel wird der IQ mit 100 multipliziert, damit er anschaulicher, weil ohne Komma geschrieben wird. In dem Beispiel: $IQ = (88/87) \cdot 100 = 101$. Der Siebenjährige erreicht einen IQ von 101.

Ein Intelligenzquotient (IQ) oder ein Entwicklungsquotient (EQ) von 1 bzw. 100 besagt: Das Intelligenz- bzw. Entwicklungsalter entspricht genau dem Lebensalter. Ein Wert über 1 bzw. 100 zeigt einen Vorsprung, ein Wert unter 1 bzw. 100 einen Rückstand an.

Probleme von Äquivalentnormen

Äquivalentnormen schließen eine Reihe von Nachteilen ein.

Es ist fraglich, ob der Quotient für jede Altersstufe das gleiche bedeutet. Gleiches Intelligenzalter bei verschiedenen Probanden kann sich aus höchst unterschiedlichen ‚Summanden‘ ergeben: Das IA von 88 Monaten etwa kann sich ergeben, wie in dem Beispiel, aus ‚6 Jahres- und 16 Monatsäquivalenten‘, aber ebenso aus ‚7 Jahres- und 3 Monatsäquivalenten‘. Behält das IA trotzdem die gleiche Bedeutung?

Darüber hinaus ist die unterschiedliche ‚Qualität‘ der Intelligenz je Altersstufe zu beachten: „Ein Neunjähriger mit dem Intelligenzalter von 6 Jahren ist eben nicht so intelligent wie ein normaler Sechsjähriger, sondern wie ein unintelligenter Neunjähriger“ (Dieterich, 1973, 190).

Äquivalentnormen lassen sich nur ermitteln bis zu einem Lebensalter von etwa 13 bis 15 Jahren. Für höhere Altersstufen finden sich keine ‚altersspezifischen‘ Testitems mehr: So gibt es ‚im Prinzip‘ keine Aufgabe, welche erst die Kohorte der 21jährigen lösen kann, dagegen ‚noch nicht‘ die Gruppe der 17jährigen. Will man für ‚höhere‘ Altersgruppen, etwa 30- oder 40jährige, Äquivalentnormen vorgeben, muß man von den Normen der jüngeren Kohorten extrapolieren.

Wichtiger, heute auch üblicher ist die zweite Normklasse.

(B) Variabilitäts- oder Abweichungsnormen

Variabilitätsnormen beruhen auf einer Transformation, bei der zu einem Rohwert ein Standardwert ermittelt wird, der angibt, wie weit der Testwert entfernt ist vom mittleren Standardwert der Eichstichprobe.

- Verteilen sich die Rohwerte normal, so daß eine lineare Transformation wiederum normalverteilte Werte erbringt, spricht man von Standardnormen.
- Verteilen sich die Rohwerte nicht normal, so kann eine Transformation die Normalverteilung herstellen. In diesem Falle spricht man von **Standard-norm-Aquivalenten**.

Interpretieren lassen sich Normen beispielsweise so, daß man drei Bereiche festlegt: einen durchschnittlichen, einen über- und einen unterdurchschnittlichen Bereich.

Als **durchschnittlich** seien etwa jene Testwerte klassifiziert, die in das Intervall ‚Mittelwert plus/minus eine Standardabweichung‘ fallen, als **überdurchschnittlich** jene Werte, die oberhalb, und als **unterdurchschnittlich** jene Werte, die unterhalb des Durchschnittsintervalls liegen.

Beispiel: Ein Test sei auf den Mittelwert 100 und die Streuung 10 standardisiert. Als **durchschnittlich** sollen Werte gelten, die in dem Intervall 100 ± 10 liegen, als **überdurchschnittlich** Werte über 110, als **unterdurchschnittlich** Werte unter 90.

Jeder empirische Test-Score ist fehlerbehaftet, darum nie punktuell zu interpretieren. Deswegen sollte man das Vertrauensintervall mitangeben (Berechnung und Problematik, S. 90 und S. 93).

Die Klassifikation nach ‚Bereichen‘ ist ‚stichprobenabhängig‘. Welche Konsequenzen sich daraus ergeben, erläutert der Abschnitt ‚Probleme der Normierung‘.

Zur Berechnung von Variabilitätsnormen

Zur linearen Transformation einer Meßwertreihe Y in eine Meßwertreihe X dient die gebräuchliche z-Formel:

$$\frac{Y - M_y}{s_y} = z = \frac{X - M_x}{s_x}$$

Es bedeuten:

Y, M_y , s_y : Einzelwert, Mittelwert, Standardabweichung der Meßwertreihe Y (etwa der Rohwerte),

X, M_x , s_x : Einzelwert, Mittelwert, Standardabweichung der Meßwertreihe X (etwa der Standardwerte).

Soll etwa ein Einzelwert Y transformiert werden in das Skalensystem von X, so ist aufzulösen nach X:

$$X = M_x + s_x \frac{Y - M_y}{s_y}$$

Mit Hilfe dieser Formel lassen sich Rohwerte in Standardwerte, Standardwerte der einen Klasse in die einer anderen Klasse umwandeln.

Beispiel: Eine Rohwertreihe Y reiche von 0 bis 20, M_y sei 12.3, s_y sei 3.9. Sie werde transformiert in Standardwerte mit $M_x = 100$ und $s_x = 10$. Für Rohwert $Y = 15$ ergibt sich dann folgender Standardwert X:

$$X = 100 + 10 \frac{15 - 12.3}{3.9} = 104.4$$

Dem Rohwert 15 (der Verteilung mit $M_y = 12.3$ und $s_y = 3.9$) entspricht der Standardwert 104.4.

Anschaulicher: Einem Wert, der in der Rohwertverteilung $,M_y + s_y'$ beträgt, also $12.3 + 3.9 = 16.2$, entspricht in der Standardwertverteilung ein Betrag von $,M_x + s_x'$, also der Wert: $100 + 10 = 110$.

(C) Prozentränge

Prozenträge (PR) beruhen auf einer Transformation, die angibt, wie groß in der Normstichprobe der Anteil von Probanden ist, deren Werte unterhalb bzw. oberhalb eines Testwertes liegen. Es handelt sich um eine reine Häufigkeitsangabe, nicht um ein Abweichungsmaß.

Beispielsweise besagt ein Prozentrang von 65: Der Testwert liegt so, daß in der Vergleichsstichprobe 65 Prozent der Probanden niedrigere oder gleiche Werte erreichen, 35 Prozent dagegen höhere Werte.

Vorgegeben seien					Berechnet werden Prozentränge (PR).						
<ul style="list-style-type: none"> - Rohwerte von 2 bis 12 (X), - Häufigkeiten je Rohwert (f), - kumulierte Häufigkeiten (cumf). 											
X	2	3	4	5	6	7	8	9	10	11	12
f	11	21	26	32	38	44	40	33	27	19	12
cumf	11	32	58	90	128	172	212	245	272	291	303
PR	4	11	29	30	42	57	70	81	90	96	100
<p>Für die Rohwertklasse X=4 berechnet sich der Prozentrang wie folgt:</p> $PR_4 = (58 \cdot 100) / 303 = 19.14$ <p>Der Maßzahlklasse 4 kommt der PR 19 zu, der besagt: Rund 19 Prozent der 303 Probanden erreichen 4 Punkte oder weniger, rund 80 Prozent erreichen mehr.</p>											

Zur Problematik transformierter Werte

Rohwerte weisen oft für unterschiedliche Gruppen verschiedene Zentralwerte auf. Kasten 4.4-1 gibt ein Beispiel: Für Neurotizismus erhalten die 14jährigen Jungen einen Mittelwert von 6.8 und eine Standardabweichung von 3.7 Rohpunkten, die 14jährigen Mädchen einen Mittelwert von 8.6 und eine Standardabweichung von 4.3.

Werden Rohpunkte für jede Gruppe isoliert in Standardwerte transformiert - etwa in Stanine-Werte - *dann wird der Unterschied der Gruppen nivelliert*, denn zufolge der Standardisierung erhält jede Gruppe denselben Mittelwert und dieselbe Standardabweichung - in dem Beispiel einen Mittelwert von 5 und eine Standardabweichung von 2 Stanine.

Wer den Unterschied zwischen Gruppen erhalten will, muß die Gruppen zu *einer* Stichprobe zusammenfassen, für Rohwerte dieser *einen* Stichprobe die Zentralwerte ermitteln und mit ihrer Hilfe die Rohwerte in Standardwerte transformieren.

In dem Beispiel muß der Untersucher die Rohpunkte der Mädchen- und der Jungengruppe *zu einer* Meßwertreihe zusammenfassen, für die Rohwerte dieser *einen* Stichprobe die Zentralwerte ermitteln, dann ihre Rohwerte (RW) in Stanine-Werte umwandeln. Das Beispiel sei ausgeführt (*M*: Mittelwert; *SD*: Standardabweichung):

$$M_{\text{gemeinsam}} = \frac{(6.8 \cdot 586) + (8.6 \cdot 524)}{(586 + 534)} = 7.74$$

$$SD_{\text{gemeinsam}} = \frac{(3.7 \cdot 586) + (4.3 \cdot 524)}{(586 + 524)} = 3.98$$

Gemeinsame Zentralwerte für Jungen und Mädchen in *Rohpunkten* (RW):

- Mittelwert: 7.74 RW,
- Standardabweichung: 3.98 RW.

Gemeinsame Stanine-Werte:

- Mittelwert: 5.00,
- Standardabweichung: 2.00.

Stanine- Werte für beide Gruppen *getrennt*:

$$\text{-- } \textbf{Jungen: } SD_J = 3.7 \frac{5 - 4.52}{7.74 - 6.8} = 1.89$$

$$\text{-- } \textbf{Mädchen: } SD_M = 4.3 \frac{5 - 5.43}{7.74 - 8.6} = 2.15$$

$$M_J = 5 + 2 \frac{6.8 - 7.74}{3.98} = 4.52$$

$$M_M = 5 + 2 \frac{8.6 - 7.74}{3.98} = 5.43$$

Jetzt übersteigen auch bei den Standardwerten, den Staninen, Mittelwert und Standardabweichung der Mädchen die der Jungen - wie bei den Rohwerten:

- M_J und SD_J betragen: $M_J = 4.52$ und $SD_J = 1.89$ Stanine,
- M_M und SD_M betragen: $M_M = 5.43$ und $SD_M = 2.15$ Stanine.

4.4.1.3 Unterschiedliche Klassifikationen von Test-Scores und übliche Normskalen

Dieses Teilkapitel verweist auf zwei Themen:

- auf unterschiedliche Klassifikationen von Test-Scores (A),
- auf übliche Normskalen (B).

(A) Verschiedene Klassifikationen von Test-Scores

Um einem Probanden seine Test-Scores verständlich zu machen, teilen wir die Werte jeweils in Klassen ein; in diesem Buch verwenden wir zur Veranschaulichung meist die Dreiteilung „durchschnittlich“, „unterdurchschnittlich“ und „überdurchschnittlich“:

- Als „**durchschnittlich**“ bezeichnen wir Werte innerhalb des Intervalles „Mittelwert ± 1 Standardabweichung“ ($M \pm 1 SD$),
- als „**überdurchschnittlich**“ Werte oberhalb von ($M + 1 SD$),
 - als „**unterdurchschnittlich**“ Werte unterhalb von ($M - 1 SD$).

Eine solche Abgrenzung kann nur *ein* Vorschlag unter mehreren sein. Andere Autoren bevorzugen es, die Klassengrenzen anders zu ziehen. Rasten 4.4-3 soll dafür Beispiele bringen.

Kasten 4.4-3: Beispiele für die Klassifikation von Test-Scores

Normalverteilung der Normwerte sei vorausgesetzt!
Welche Intervalle kommen in Betracht?

Vier Beispiele für die Festlegung von „Durchschnittsintervallen“:

1. Mittelwert ± 1 Standardabweichung:

So in der Regel das Vorgehen in diesem Buch, etwa in dem Beispiel zu Kasten 4.4-1 (S. 112)!

2. Mittelwert ± 0.5 Standardabweichung:

Zwei Beispiele: (1) „16 Personality Factor Questionnaire“ (Cattell et al., 1970), (2) „Freiburger Persönlichkeitsinventar“ (Fahrenberg et al., 1989).

3. Mittelwert ± 1 Standardmeßfehler:

Beispiel: Wechsler-Test-Serie (Wechsler, 1964; Tewes, 1991).

Konsequenz: Die Spanne des Durchschnittsbereiches variiert mit der Höhe der Reliabilität und der Spanne der Standardabweichung.

4. Median ± 1 Quartil:

Der Durchschnittsbereich umfaßt die Prozentränge 25-75. Diese Klassifikation wird in der Regel gewählt bei Test-Scores, für die nur Prozentränge zur Verfügung stehen. - Drei Beispiele: (1) Der „Picture Frustration Test“ (Rosenzweig, 1945, 1957), (2) die „Sorgfaltsleistung“ (F %) im „Aufmerksamkeits-Belastungstest d2“ (Brickenkamp, 1994) (3) die KURZFORM des „Freiburger Persönlichkeitsinventars (FPI-K)“ (Fahrenberg et al., 1989).

(B) Übliche Normskalen

In der Testkonstruktion haben sich unterschiedliche Normskalen eingebürgert. Je mehr Einheiten sie enthalten, desto feiner - so suggerieren sie - differenzieren sie. Doch hängt die Differenzierungskraft, wie das Konzept der ‚Vertrauensintervalle‘ zeigt, von der Reliabilität eines Test ab.

Einige **Normskalen** seien in Kasten 4.4-4 mit Mittelwert und Streuung angeführt sowie mit einem Bereich von ‚Mittelwert drei Standardabweichungen‘.

„Es gibt keinen vernünftigen Grund, warum so viele Normskalen im Gebrauch sind. Man könnte sich eigentlich einmal auf ein paar vernünftige Skalen einigen. Dabei könnte man einen oder zwei Mittelwerte festlegen und einige verschiedene Streuungen, so daß man fein differenzierende und grob differenzierende Tests konstruieren kann“ (Dieterich, 1973, 194).

Kasten 4.4-4:
Einige übliche Normskalen

Der ‚Bereich‘ gibt die Spanne ‚Mittelwert ± 3 Standardabweichungen‘ an.			
Skala	Mittelwert	Standardabweichung	Bereich
z-Werte	0	1	-3 bis +3
IQ	100	15	55 bis 145
Z-Werte	100	10	70 bis 130
T-Werte	50	10	20 bis 80
Centile	5	2	-1 bis 11
Stanine*)	5	2	+1 bis 9

*) Stanine, ein Kürzel für ‚Standard Nine‘, reichen vom Wert 1 bis zum Wert 9.

HINWEIS: Das Verhältnis verschiedener Normskalen zueinander veranschaulichen Lienert und Raatz in „Tafel 2“ ihres Anhanges unter dem Titel „Die Transformation von Testnormen“ (1994, 410).

4.4.2 Probleme der Normierung

Normen zu erstellen ist eine statistische Prozedur, - die jedoch übergreift in die Problematik anderer Normen. Einige Probleme seien genannt:

- Wahl der Eichstichproben (4.4.2.1),
- Normalverteilung der Merkmalsausprägungen (4.4.2.2),
- Normbezogene Klassifikation und Stichprobenabhängigkeit (4.4.2.3),
- Statistische Normen und kultureller Kontext (4.4.2.4).

4.4.2.1 Wahl der Eichstichproben

Die Auswahl von Eichstichproben stellt den Testautor vor die Frage: Wie kann er verbürgen, daß jeder Proband, der zur angezielten Population gehört, eine berechenbare und von Null verschiedene Chance habe, auch in die Eichstichprobe einzugehen?

Zwei Modelle helfen die Frage zu lösen: Ziehung von Flächen- oder von Quotenstichproben.

Flächenstichprobe: In einem bestimmten Bezirk wird nach einem Schlüssel eine Stichprobe gezogen: nach System (etwa jeder hundertste Bewohner) oder nach Zufall (etwa nach Zufallszahlen).

Quotenstichprobe: Die Struktur der Eichstichprobe wird bestimmt, z.B. nach Determinanten wie Alter, Geschlecht, Beruf, Wohngebiet. Dann werden Quoten von Probanden ermittelt, so daß die relevanten Teilgruppen in der Stichprobe 'abgebildet' werden, d.h. gleiche Anteile wie in der Population aufweisen.

Ein Sonderproblem der Selektion besteht darin, die Eichstichprobe so zusammenzusetzen, daß sie die gleiche Struktur hat wie die Population, auf die verallgemeinert werden soll. (Sonst lassen sich die bereits ermittelten Kennwerte nicht übertragen.)

Dieses Problem wiederholt sich, wenn innerhalb der Gesamtgruppe **Teilstichproben** gebildet werden, zum Beispiel nach Alter, nach Geschlecht, nach Bildungsstand. Jedesmal ist erneut zu prüfen, wie weit die Kennwerte (für Item und Test) in Teil- und Gesamtstichprobe gleich bleiben.

4.4.2.2 Normalverteilte Merkmalsausprägung als Voraussetzung der Normierung

Wenn Roh- und Standardwerte Aussagen über das Ziel-Merkmal ermöglichen sollen, setzen sie voraus, daß die Ausprägungen des Ziel-Merkmals sich normal verteilen.

Das Problem sei an dem Bild von Kopie und Original veranschaulicht: Bei den Items und ihren Scores handelt es sich um Kopien, nicht um Originale. Die ‚Originale‘ sind die latenten Ziel-Merkmale - Konstrukte, die nicht beobachtbar sind. Die Test-Scores können nur als ‚Kopien‘ gelten, wenn sie im Beobachtbaren die Struktur der Originale nachbilden. Normalverteilte Merkmale lassen sich nur in normalverteilten Test-Scores abbilden.

Ob die latenten Merkmale normalverteilt sind, ist keine Frage, die sich empirisch beantworten läßt; sie ist nur theoretisch zu klären. „Man kennt sehr gut die Bedingungen, unter denen sich die Ausprägungen von Merkmalen normal verteilen:

1. Bedingung: Die individuelle Ausprägung eines Merkmals muß von zahlreichen Faktoren abhängen.
2. Bedingung: Diese Faktoren müssen unabhängig voneinander sein.
3. Bedingung: Die Merkmalsausprägung muß durch additives Aufeinanderwirken dieser Faktoren zustandekommen“ (Dieterich, 1973, 188).

Ist es gewiß, daß solche Bedingungen bei psychischen Merkmalen erfüllt sind?

4.4.2.3 Normbezogene Klassifikation und Stichprobenabhängigkeit

Die Klassifikation nach ‚Normbereichen‘ ist ‚stichprobenabhängig‘. Sie interpretiert die Test-Leistung in Relation zur Eichstichprobe.

Beispiel: *Eine Intelligenztest-Leistung, für die 107 SW gegeben werden, wird von uns als ‚durchschnittlich‘ klassifiziert im Vergleich zu den in der Eichstichprobe ermittelten Werten.*

Anders verhält es sich bei ‚absoluten‘ Maßen wie Länge oder Schwere. Über die Größe eines Kindes lassen sich beispielsweise zwei Arten von Aussagen machen:

- Erstens läßt sich die ‚absolute‘ Feststellung treffen, das Kind sei 107 cm groß. Da die Einheit Zentimeter ‚absolut‘ festgelegt ist, bildet diese Maßangabe eine verständliche Aussage, die keinen zusätzlichen Kontext braucht.
- Zweitens läßt sich sagen, das Kind sei ‚groß im Vergleich‘ zu anderen Kindern seines Alters. Die ‚absolute Feststellung‘ läßt sich also in eine ‚relative Angabe‘ umwandeln: in eine stichprobenabhängige Aussage.

Beide Aussagen sind sinnvoll, aber jede steht in einem anderen Kontext: die zweite ist ‚veränderlicher‘ - sie ändert sich, wenn sich die Bezugsgruppe ändert.

Testwerte sind nach der zweiten Aussagenart zu interpretieren, sie liefern ‚nur‘ Vergleichsaussagen, sie sind stichprobenabhängig. Sie sind begrenzt

- auf die Kultur,
- auf die Zeit,

- auf die Region, der die Eichstichprobe angehört.

Beispiel: Der ‚Hamburg-Wechsler-Intelligenztest für Erwachsene‘ (HA WIE) wurde in den 50er Jahren geeicht (Wechsler, 1964). Somit liefern seine Normen Vergleiche zu ‚Leistungen‘ der 50er Jahre. Wie weit rechtfertigt sich ein solcher Vergleich noch nach zwanzig, dreißig oder vierzig Jahren? - Aufgrund von Fragen wie diesen wurde der HA WIE Ende der 80er Jahre völlig überarbeitet und neu normiert. Auf diese Weise wurde seine ‚epochale‘ Stichprobenabhängigkeit anerkannt, der Test wurde an ‚heutige‘ Bedingungen adaptiert (Tewes, 1991; vgl. die Bewertung von Fay, 1993, in diesem Buch S. 291).

4.4.2.4 Statistische Normen und kulturell-ethischer Kontext

Der statistischen Normierung liegen (auch) Wertungen und Festlegungen zugrunde, die nicht statistischer Natur sind.

Das sei verdeutlicht: Ein Intelligenztest möge Aufgaben enthalten, die den Umfang des aktiven Wortschatzes oder den Grad der Abstraktionsfähigkeit erfassen sollen.

Wer legt fest, daß ein ‚großer Wortschatz‘ oder eine ‚hohe Abstraktionsfähigkeit‘ wichtige kognitive Funktionen sind? Der Testautor nicht allein, sondern mit ihm der Testanwender, ebenso der Proband, der getestet wird, kurz: die ‚Gesellschaft‘, zu der sie alle gehören. In diesem Sinne beruht das statistische Urteil auch auf sozio-kulturellen Urteilen, die der Testautor mit anderen Menschen seiner Kultur teilt.

Die ‚Gesellschaft‘ besteht aus Menschen, die in sozialen Schichten, in einer bestimmten Epoche, in einer bestimmten Kultur Maßstäbe gesetzt haben. Solche Maßstäbe legen mit fest, was in Tests als ‚wichtig‘ gilt.

Insofern betreffen statistische Normen Werte unterschiedlicher Art. Statistische Normen heben an einem Test-Merkmal nur einen Aspekt hervor: den der Häufigkeitsverteilung in einer Population. Die anderen Aspekte, unter denen dasselbe Merkmal betrachtet und bewertet wird, lassen sich dabei zwar ausblenden, behalten aber ihre Bedeutung.

4.5 Beitrag zu Diagnostik und Intervention

Was Kapitel 9 und 10 im einzelnen belegen, sei hier nur skizziert (S. 270 und 313). Die klassische Testtheorie eignet sich mehr zur Diagnostik als zur Intervention.

Ein Beitrag zur Diagnostik läßt sich unter folgenden Perspektiven benennen:

- Tests ermöglichen eine wohldefinierte Merkmalserfassung - dank Konstruktionsregeln und Anwendungsvorschriften.
Somit beschreiben Tests - *im Idealfall* - ein Merkmal in einem theoretischen Kontext.
Tests ermöglichen eine Vielzahl von Vergleichen:
⇒ Vergleiche von *Individuen* mit Individuen oder mit Gruppen,
⇒ Vergleiche von *Gruppen* mit Individuen oder mit Gruppen.
Sie ermöglichen somit Klassifikationen für weitere psychologische Prozeduren.
Tests tragen zur Sprachregelung zwischen den Diagnostikern bei.

Ein Beitrag zur Intervention könnte in folgenden Anteilen liegen:

- Tests können helfen, Interventionsbedarf festzustellen.
- Tests oder testähnliche Verfahren können der Prävention und dem Training dienen.
- Tests können die Bilanzierung einer Intervention erleichtern.

4.6 Kritik der klassischen Testtheorie

Tests, konstruiert nach der klassischen Testtheorie, haben sich bewährt (Michel & Conrad, 1982, 25). Doch sollte, wer sie anwendet, ihre Schwachstellen kennen und bei der Verwendung beachten (Fischer, 1974, 114-137; Michel & Conrad, 1982, 24-26). Dazu einige Hinweise, die nur zusammenfassen, was die vorausgehende Darstellung vereinzelt schon gesagt hat.

- a) Die klassische Testtheorie beruht auf einer **Fehlertheorie**, die nicht psychologisch fundiert ist. Die Axiome sind nicht abgeleitet aus einer psychologischen Reflexion oder einer psychologischen Theorie.
- b) Die klassische Testtheorie setzt Daten auf dem Niveau einer **Intervallskala** voraus. Denn es werden Mittelwerte und Varianzen berechnet, es werden Differenzen von Meßwerten gebildet. Aber es ist fraglich, ob Testdaten das Meßniveau von Intervallskalen erreichen.
- c) **Der wahre Wert** (als Repräsentant einer latenten Fähigkeit) wird, bezogen auf eine Person, als **invariant** betrachtet. Doch kommen der psychischen Realität Annahmen näher, die besagen, daß die ‚wahren Fähigkeiten‘ einer Person fluktuieren.
Dann aber wird auch die Annahme einer Unkorreliertheit von Fehlerwert und wahren Wert fragwürdig. Es läßt sich vorstellen, daß in besonderen Testbereichen Fehler und wahrer Wert systematische Zusammenhänge eingehen.
- d) **Folgerungen** aus der klassischen Testtheorie erweisen sich als **problematisch**: Nach dem Verdünnungsparadox (attenuation paradox) sinkt die kriterienbezogene Validität eines Verfahrens mit wachsender Reliabilität von Kriterium und validiertem Test.

e) Alle Meßwerte der klassischen Testtheorie sind **stichproben- oder populationsabhängig**. Darum bleibt es immer schwierig zu bestimmen, wie weit sie generalisierbar sind.

Diese Abhängigkeit bringt es mit sich, daß zum Beispiel die Reliabilität künstlich erhöht oder gesenkt werden kann:

- erhöht dadurch, daß ein Verfahren einer heterogenen Gruppe vorgelegt wird (heterogen bezüglich des zu erfassenden Merkmals),
- gesenkt dadurch, daß sie einer homogenen Stichprobe vorgelegt wird (wieder homogen bezüglich des zu erfassenden Merkmals).

Populationsabhängige Aussagen bleiben sinnvoll, sind aber in ihrem Umfang regional und epochal begrenzt.

f) Weil die Meßwerte der klassischen Testtheorie an Stichproben gewonnen wurden, ist **in bestimmten Fällen die Anwendung auf individuelle Fälle problematisch**, zum Beispiel wenn Gruppenstatistiken auf Individuen übertragen werden, etwa bei der Berechnung von Vertrauensbereichen.

Überblick zu Test-Kennwerten

Den Abriß der klassischen Testtheorie schließe Kasten 4.6-1 mit einer Übersicht über Bereiche wichtiger Testkennwerte.

Kasten 4.6-1: Beurteilung der Höhe von Testkennwerten

(nach Weise. 1975, 210)

<i>Kennwert</i>		<i>Kürzel</i>	<i>Niedrig</i>	<i>Mittel</i>	<i>Hoch</i>
Schwierigkeit		p	> .80	.80- .20	< .20
Trennschärfe	(korrigiert)	r_{ite}	< .30	.30- .50	> .50
Objektivität	(Auswerter)	r_k	< .60	.70- .90	> .90
Reliabilität		r_{tt}	< .80	.80- .90	> .90
Validität*)	(unkorrigiert)	r_{ic}	< .40	.40- .60	> .60
Eichsstichprobe		N	<1.50	150-300	>300

*) Validitätskoeffizienten sagen wenig über die Bedeutung eines Tests, wenn man nur den absoluten Betrag bewertet. Beachtet werden muß auch der Beitrag, den ein Test zur Lösung einer gegebenen Fragestellung leisten kann (vgl. Basis-, Selektionsrate und Validität, S. 346, Nutzenschätzung, S. 373)

4.7 Zusammenfassung zu Kapitel 4

Kapitel 4 hat einige Grundgedanken der klassischen Testtheorie vorgestellt, indem es die Genese eines Tests verfolgte.

Bei der **Fragestellung** geht es darum, das Testmerkmal scharf abzugrenzen - im Idealfall das Merkmal *theoriegeleitet* zu definieren.

Welche Items zu einem Test zusammengefaßt werden, läßt sich entscheiden nach drei **Konstruktionsstrategien**: *Items werden zu einer Skala zusammengestellt*

- bei *rationaler Strategie*, wenn sie einem vorgegebenen theoretischen Konzept entsprechen;
- bei *externaler Strategie*, wenn sie zwischen einer Kriteriums- und einer Kontrollgruppe unterscheiden;
- bei *internaler Strategie*, wenn eine statistische Prozedur sie als zusammengehörig erweist.

Bei der **Item-Generierung** ist es schwierig, die Items streng theoretisch abzuleiten und regelhaft zu ‚vervielfältigen‘.

Items lassen sich **unterscheiden** unter verschiedenen Aspekten; genannt seien drei:

- Aufgaben nach Antwortart: *gebunden oder frei*,
- Aufgaben nach Inhaltsumfang: *einfach oder komplex*,
- Aufgaben nach Darstellungsmedium: *verbal oder nichtverbal*.

Die **Itemanalyse** soll mit statistischen Kennwerten die ‚geeignetsten Items‘ identifizieren. Ermittelt werden vor allem drei Kennwerte:

- Die *Trennschärfe* gibt an, wie hoch die Korrelation zwischen Item-Score und Test-Score ausfällt.
- Der *Schwierigkeitsindex* gibt an, wie groß der Anteil der Löser oder der Lösungen an der maximal erreichbaren Summe der Wertepunkte ist.
- Die *Homogenität* gibt an, wie hoch ein einzelnes Item mit den anderen Items korreliert.

Bei der **Itemselektion** ist zu entscheiden, welche Items zu behalten und welche zu eliminieren sind:

- Für *homogene* Tests läßt sich als Hilfe ein statistisches Kriterium ermitteln, der *Selektionskennwert*: ein Index, der für die Entscheidung über die Items Schwierigkeit und Trennschärfe verbindet.
- Bei *heterogenen* Tests werden sukzessive Einzelschritte vorgeschlagen: Ausgeschieden werden Items, deren Schwierigkeit und Trennschärfe einem vorgegebenen Kriterium nicht genügen. Von den verbliebenen Items behält man jene mit höherer Trennschärfe.

Drei **Gütekriterien** bestimmen, wie angemessen der Gesamttest das Testmerkmal abbildet:

- a) **Objektivität** bezeichnet das Maß, wie weit in der diagnostischen Situation der gesamte Testvorgang standardisiert wird. Man unterscheidet: Durchführungs-, Auswertungs- und Interpretationsobjektivität.
- b) **Reliabilität** umschreibt die Meßgenauigkeit eines Tests. Zurückgeführt wird sie auf sogenannte Axiome, die auf der Annahme beruhen, daß man in einem beobachteten Meßwert zwei Anteile unterscheiden kann: wahren Wert und Fehler.

Geschätzt werden die Anteile von wahrem Wert und Fehler nach vier Paradigmen: *Retest*-, *Paralleltest*- und *Halbierungsreliabilität* sowie *Konsistenz*.

- c) **Validität** bezeichnet die Beziehung zwischen Test-Score und Test-Merkmal. Man unterscheidet drei Arten:
- *Inhaltsvalidität* besteht darin, daß der Inhalt der Test-Items sich laut Experten-Urteil selber rechtfertigt.
 - *Kriterienbezogene Validität* besteht darin, daß der Test-Scores mit einem Kriteriums-Scores korreliert wird: (1) Wird der Kriteriums-Score gleichzeitig mit dem Test-Score ermittelt, so spricht man von Übereinstimmungsvalidität. - (2) Wird der Kriteriums-Score später ermittelt als der Test-Score, so spricht man von Vorhersagevalidität.
 - *Konstruktvalidität* vereinigt Schritte der inhaltlichen und kriterienbezogenen Validität. Sie besteht darin, daß ein Testmerkmal in ein nomologisches Netz von Aussagen eingebunden wird.

Als Beispiel für die Realisierung der Konstruktvalidierung läßt sich die **Multitrait-Multimethod-Validierung** betrachten. Mehrere Merkmale werden mit mehreren Methoden erfaßt, so daß es möglich wird, die Varianzanteile der Methoden und der Merkmale zu trennen.

Normen ermöglichen es, anzugeben, welche Position ein Proband einnimmt bezüglich der Werte einer Normstichprobe:

- a) **Rohwerte** bestehen in der Zahl der Punkte, die für „Antworten in Schlüsselrichtung“ gegeben werden. - Ein Nachteil von Rohpunkten liegt darin, daß verschiedene Tests unterschiedliche Zentralwerte aufweisen können und dadurch Vergleiche erschweren.
- b) **Transformierte Werte** ergeben sich, wenn Rohwerte in gleiche Normskalen „umgewandelt“ werden. Man unterscheidet drei Hauptklassen:
- *Äquivalentnormen* beruhen auf einer Transformation, bei der einem Rohwert ein Zeit-Äquivalent zugeordnet wird.
 - *Variabilitätsnormen* beruhen auf einer Transformation, bei der zu einem Rohwert ein Standardwert ermittelt wird, der angibt, wie weit ein einzelner Testwert entfernt ist vom mittleren Standardwert der Normstichprobe.
 - *Prozentränge* beruhen auf einer Transformation, bei der ermittelt wird, wie groß in der Normstichprobe der Anteil von Probanden ist, deren Werte unterhalb und oberhalb eines einzelnen Testwertes liegen.

An der klassischen Testtheorie wird unter verschiedenen Aspekten **Kritik** geübt; einige dieser Punkte besagen:

- Die klassische Testtheorie beruht auf einer *Fehlertheorie*, die nicht psychologisch fundiert ist.
- Die klassische Testtheorie setzt Daten auf dem Niveau einer *Intervallskala* voraus. Aber es ist fraglich, ob Testdaten dieses Meßniveau erreichen.

- Der wahre Wert wird, bezogen auf eine Person, als *invariant* betrachtet. Doch ist eher anzunehmen, daß die ‚wahren Fähigkeiten‘ einer Person fluktuieren.
- Alle Meßwerte der klassischen Testtheorie sind *stichproben- oder populationsabhängig*. Darum bleibt es immer schwierig zu bestimmen, wie weit sie generalisierbar sind.

4.8 Kontrollfragen zu Kapitel 4 (*KTT: Klassische Testtheorie*)

Umschreibung der KTT.

Axiome der KTT.

Bestimmungsstücke von ‚Test‘ im Sinne der KTT.

Bedeutung der Fragestellung und der Merkmalsabgrenzung für eine Testkonstruktion.

Drei Strategien einer Testkonstruktion.

Gesichtspunkte der Generierung von Items.

Gliederung von Items nach ihren ‚Bauformen‘.

Schritte und Ziele der Itemanalyse.

Vier Konzepte von Homogenität.

Objektivität: Definition, Arten.

Reliabilität: Definition, Arten.

Standardmeßfehler eines Test-Scores.

Kritische Differenz zwischen Test-Scores.

Validität: Definition, Arten.

Konstruktvalidierung.

Multitrait-Multimethod-Validierung.

Normierung.

Einwände gegen die KTT.

5. Kapitel

Hinweise zur kriteriumsorientierten Leistungsmessung

Neben und aus der klassischen Testtheorie wurde ein Meßkonzept entwickelt, nach dem die Leistung eines Probanden nicht verglichen wird mit stichprobenbezogenen Normen, sondern mit einem inhaltlich definierten Ziel. Das Ziel heißt ‚Kriterium‘. Darum spricht man von kriteriumsorientierter Leistungsmessung.

Anschaulich und sehr differenziert hat Klauer diesen Ansatz dargestellt unter dem Titel „kriteriumsorientierter Tests“ (1987). An diesem Werk orientiert sich zu wesentlichen Teilen das folgende Kapitel.

Die Probleme, die wir besprechen, gliedern wir in fünf Abschnitte:

- Abgrenzungen: kriteriumsorientierte Leistungsmessung und kriteriumsorientierter Test (5.1),
- Konstruktion kriteriumsorientierter Testaufgaben (5.2),
- Analyse kriteriumsorientierter Testaufgaben (5.3),
- Schluß vom Test-Score auf die Fähigkeit eines Probanden (5.4),
- Beitrag zu Diagnostik und Intervention (5.5).

Das Kapitel schließt mit einer Zusammenfassung (5.6) und der Vorgabe einiger Kontrollfragen (5.7).

5.1 Abgrenzungen: kriteriumsorientierte Leistungsmessung und kriteriumsorientierter Test

Kriteriumsbezogene Tests lassen erkennen, ob ein Proband Aufgaben lösen kann, die ein Kriterium umschreibt. Ein **lehrziel-orientierter** Test soll darüber informieren, ob ein ‚Schüler‘ ein vorher festgelegtes Lehrziel, ein **therapieziel-orientierter** Test soll prüfen, ob ein ‚Klient‘ ein vorher festgelegtes Therapieziel erreicht hat. „Diese Tests sind also idealnormiert und müssen eine hohe Inhaltsvalidität aufweisen, da die Aufgaben Stichproben des Zielverhaltens sein sollen“ (Leichner, 1979, 115).

Kriteriumsbezogene Tests **zielen**

- **nicht** auf Erfassung individueller Differenzen (wie die klassischen Tests), also nicht auf Ermittlung des Rangplatzes eines Probanden in einer vergleichbaren Personengruppe (Normgruppe, Population),
- **sondern** auf Feststellung der Leistung eines Probanden bezüglich eines spezifizierten Aufgabenbereiches (z. B. eines Therapie- oder Lehrzieles).

Demgemäß unterscheidet sich ein ‚kriteriumsbezogener‘ von einem ‚normbezogenen‘ Test der klassischen Testtheorie vor allem dadurch, daß er die Leistungen eines Probanden nicht mit empirisch festgestellten Durchschnittsleistungen einer Eichstichprobe, sondern mit vorher festgelegten Inhalten vergleicht, die das Kriterium repräsentieren.

So definiert sich ein kriteriumsorientierter Test durch seine inhaltliche Validität (Klauer, 1987, 11).

„Ganz allgemein kann er als ein Test zur Erfassung eines Persönlichkeitsmerkmals verstanden werden, sofern das Persönlichkeitsmerkmal durch die ‚Bewältigung‘ einer wohldefinierten Aufgabenmenge gekennzeichnet ist.“

Für Lehr- und für Therapieziele gilt **eine weiter entfaltete Umschreibung** (Klauer, 1987, 11):

„Kriteriumsorientiert ist ein Test, der die Gesamtheit einer wohldefinierten Menge von Aufgaben enthält oder repräsentiert und der zu dem Zweck konstruiert ist,

- *die Fähigkeit des Probanden zur Lösung der Aufgaben der definierten Menge zu schätzen und/oder*
- *ihn gemäß dieser Fähigkeit einer Klasse von Probanden zuzuordnen“.*

Beispiele: Zur Veranschaulichung sei auf zwei bekannte kriteriumsorientierte Tests verwiesen:

1. *auf die schriftliche Prüfung zur Erlangung des Führerscheins,*
2. *auf die staatlichen Medizinerprüfungen. Um zu bestehen, muß der Prüfling in beiden Fällen einen bestimmten Punktwert erreichen, der seinerseits inhaltliche Kriterien repräsentiert.*

Ein Gegensatz: kriteriumsorientiert versus normorientiert?

Gelegentlich kontrastiert man ‚kriteriumsorientierte‘ und ‚normbezogenene‘ Leistungsmessung sehr scharf. Doch ist weder der Bedeutungshof von ‚Kriterium‘ noch der von ‚Norm‘ eindeutig umrissen, der Kontrast darum auch nicht prägnant.

Denn in diesem Kontext kann ‚Kriterium‘ mindestens drei verschiedene Bedeutungen annehmen: Es kann, *erstens*, ein Lehrziel bezeichnen, das erreicht oder nicht erreicht wird. Es kann, *zweitens*, ein Leistungskontinuum umschrei-

ben, auf dem unterschiedlich ‚tüchtige‘ Probanden unterschiedliche Positionen einnehmen. Es kann, *drittens*, für einen Leistungsstandard stehen, an dem sich Vorhersagen bestätigen (oder widerlegen) lassen (Klauer, 1987, 3).

Auch ‚Norm‘ kann unterschiedliche Sachverhalte charakterisieren: Sie kann sich auf empirische Kennwerte einer Bezugsgruppe, z.B. auf Eichwerte im Sinne der klassischen Testtheorie, also auf eine Realnorm beziehen. Sie kann aber auch einen Kanon von Anforderungen, z.B. inhaltliche Lehrziele einer Ausbildung, also eine Idealnorm umschreiben (Klauer, 1987, 4).

5.2 Konstruktion kriteriumsorientierter Testaufgaben

Den kriteriumsorientierten Test bestimmt der Inhalt des Kriteriums. Darum muß ein Testautor angeben, wie er das Kriterium definiert und wie er die ‚Inhalte‘ des Kriteriums in den Item-Mengen ‚abbildet‘. Nun gibt es „zwei Möglichkeiten, Mengen zu definieren, die vollständige Aufzählung aller ihrer Elemente und die Angabe mengenbildender Merkmale“ (Klauer, 1984, 5; vgl. Fricke, 1973, 115; Jackson, R. 1973, 92-102; Klauer, 1987, 12-53).

Eine ‚enumerative Aufzählung‘ der Testaufgaben als der Elemente, die das Kriterium vollständig umschreiben, dürfte die Ausnahme bleiben; der Bedeutungshof der meisten Kriterien dürfte zu groß sein. Darum kommt es in der Regel darauf an, ‚mengenbildende Merkmale‘ von Kriterien und von Testaufgaben präzise zu beschreiben.

Beispiele für das Vorgehen seien übernommen von Fricke (1974, 23-37) und von Klauer (1987, 17-28):

- Operationale Definition (5.2.1),
- Aufspaltung der Aufgaben nach Zielen und Inhalten (5.2.2),
- Generative Regeln (5.2.3).

5.2.1 Operationale Definition

Das Kriterium wird definiert, indem repräsentative Inhalte operational beschrieben werden. So kann etwa festgelegt werden: Ein pädagogischer Test soll ‚Rechnerisches Denken‘ prüfen mit Aufgaben aus der Zinsrechnung. Die Testaufgaben sollen vier Größen enthalten: Zinsertrag, Kapital, Zinsfaktor und Jahre der Verzinsung. Kasten 5-1 gibt ein Beispiel.

Kasten 5-1:

Generierung von Testaufgaben nach operationaler Definition (Fricke, 1974, 23)

Es wird folgende Zinsformel vorgegeben:

$$\text{Zinserertrag} = \frac{\text{Kapital} \cdot \text{Zinsfaktor} \cdot \text{Jahre der Verzinsung}}{100}$$

Aufgabe: Löse die Zinsformel nach jeder der drei Größen auf der rechten Seite der Gleichung auf!

Solche Aufgaben sollen drei **Kriterien** genügen (Fricke, 1974, 23):

1. *Beschreibung des geforderten Verhaltens*, hier: Die Zinsformel soll nach jeder Größe auf der rechten Seite aufgelöst werden.
2. *Konkrete Bedingungen*, hier: Der Proband darf schriftlich arbeiten.
3. *Beurteilungsmaß*, hier: Zehn Aufgaben dieser Art sollen in dreißig Minuten gelöst werden.

5.2.2 Aufspaltung der Aufgabe nach Zielen und Inhalten

Ein therapeutisches Design könnte in einer Matrixform dargestellt werden nach den Zielen, die erreicht werden sollen, und nach den Inhalten, die das Ziel ausmachen (sogenannte Tyler-Matrix).

Bei einem Klienten liege eine Schlangenphobie vor. Die *Inhalte* des therapeutischen Vorgehens sollen die Art der Schlangendarstellung betreffen, die *Ziele* den Grad der Annäherung an die ‚Inhalte‘. Kasten 5-2 gibt ein Beispiel.

Kasten 5-2:

Aufgabenmatrix zur Quantifizierung eines Therapiezieles (Fricke, 1974, 25)

Es liege eine Schlangenphobie vor. <i>Ziele</i> und Inhalte einer Therapie werden festgelegt.				
<i>Inhalte</i>	<i>Ziele:</i> Annäherung an die ‚Schlange‘			
	<i>Auf 5 m</i>	Auf 2 m	Auf 1 m	Berühren
Schlangenbild				
Plastikschlange				
Ausgestopfte Schlange				
Lebende Schlange				

Der Erfolg kann in den Zellen angekreuzt, in diesem Sinne auch quantifiziert werden.

5.2.3 Generative Regeln

Es wird ein **Sachverhalt** vorgegeben und eine **Aufgabenform** gewählt (z.B. Ergänzungsaufgaben: Das Problem wird vorgegeben, der Proband muß die

Frage *kurz* ergänzen). **Transformationsregeln** legen fest, wie der Sachverhalt in die *Aufgabenform* übertragen wird. Kasten 5-3 gibt ein Beispiel.

Kasten 5-3:
Erzeugung von Testaufgaben mittels generativer Regeln
(Klauer, 1987, 22)

Leitidee:	Es wird geprüft, wie weit die vier Grundrechenarten im Zahlenraum von 1 bis 100 (N_{100}) beherrscht werden.
Sachverhalt:	$\Rightarrow a \circ b = c$ $\Rightarrow a, b, c \in N_{100}$ (Menge der natürlichen Zahlen von 1 bis 100) $\Rightarrow \circ \in \{+, -, \cdot, : \}$
Itemform:	Ergänzungsaufgabe
Transformationsregeln:	<ol style="list-style-type: none"> 1. Ersetze die vier Variablen der Aussageform $a + b = c$ durch Elemente von N_{100} so, daß eine ‚wahre Aussage‘ entsteht. 2. Streiche dann a oder b oder c.

Auf diese Weise kann eine Reihe wohldefinierter Items entworfen werden. „Die Gesamtheit der Aufgaben, die den Transformationsregeln entsprechend erzeugt werden kann, ist die **Grundmenge** von Aufgaben“ (Klauer, 1987, 24; vgl. Feger, B. 1984).

Grenzen der Konstruktion von Aufgaben

Die Beispiele bei Klauer (1987) zeigen ein hohes Maß an didaktischem Geschick, einen Lehrstoff in Anteile zu zerlegen, die sich für Einzelaufgaben eignen.

Doch werden auch Grenzen sichtbar: Nicht jeder Lehrstoff läßt sich schlüssig in ‚Teile‘ aufgliedern, die dann Items abgeben. Am besten eignen sich Sachgebiete, deren Teilbereiche klar und eindeutig voneinander abzugrenzen sind: etwa Einheiten aus Mathematik, Physik oder Geographie.

Weniger prägnant lassen sich Einheiten abgrenzen, wenn es um Zusammenhänge geht, die sich vielfältig verschränken und überschneiden, etwa geschichtliche Verläufe oder sprachliche Gestaltungen.

Kasten 5-4 führt ein Beispiel an, das sowohl Klauers Geschick bei der Zerlegung des Lehrstoffes demonstriert, aber auch die Grenzen anzeigt, auf die eine solche Zerlegung stößt.

Kasten 5-4:**Erzeugung von Testaufgaben: Sachverhalte darstellen in Aussagenmengen**

(Klauer, 1987, 34)

Beispiel: Ursachen, Verlauf und Folgen des Dreißigjährigen Krieges**Sachverhalt:**

Der Sachverhalt ist in einem längeren Lehrbuchtext dargestellt. Er wird der Einfachheit halber hier nicht wiederholt.

Anmerkung: Meistens ist es günstiger, den Sachverhalt neu darzustellen. Der Ausgangstext sollte nichts Überflüssiges, keine Weitschweifigkeiten sowie keine Wiederholungen enthalten, auf der anderen Seite aber vollständig sein, d.h. alles ausformuliert enthalten, was sonst vielfach nur implizit mitgemeint ist, und er sollte sprachlich angemessen sein.

Aufgabenform:

Stimulus-Komponente: W-Frage (Wer? Was? Wann? Wo?)

Reaktions-Komponente: Richtige Antwort

Transformationsregeln:

- 1) Wähle für jeden der Aussagesätze nach Zufall ein Satzglied
- 2) Formuliere eine W-Frage nach diesem Satzglied

Zerlegung in Teilmengen:

Der Text wird in fünf Absätze zerlegt, die folgende Bezeichnungen erhalten:

- Ursachen des Krieges,
- Anlaß und Ausbruch des Krieges,
- Verlauf,
- Friedensbemühungen und Friedensschluß,
- Folgen.

Samplingvorschriften:

- 1) Die fünf Teilmengen von Sätzen sind gleich stark zu berücksichtigen.
- 2) Innerhalb der Teilmengen erfolgt eine Zufallsauswahl der Sätze (Alternative: Der Testautor wählt die seiner Meinung nach wichtigsten Sätze aus).
- 3) Gruppe A der Fragewörter (wer, was, wann, wo) und Gruppe B (aus welchem Grund, zu welchem Zweck) werden im Verhältnis A:B = 2:1 eingesetzt.

Kurzfassung:

Gegeben werden Fragen zu Ursachen, Verlauf und Folgen des Dreißigjährigen Krieges. Der Schüler beantwortet die Fragen mit eigenen Worten.

Die Schwierigkeiten, die sich einem Item-Konstrukteur stellen, seien in Fragen zu Kasten 5-4 gefaßt:

- Läßt sich der *Sachverhalt* so eindeutig eingrenzen, daß wichtige Teile weder wiederholt noch ausgelassen werden? (Es geht um die Frage der Repräsentativität, die frei von Redundanz ist.)
- Ist bei einer so komplexen Materie zu erwarten, daß sich objektive Abgrenzungen von fünf *Teilmengen* ergeben? (‚Objektiv‘ sei verstanden in dem Sinne, daß sich verschiedene Konstrukteure auf gleiche Einheiten einigen.)
- Sind die erzeugten Fragen *gleichgewichtig*? (‚Gleichgewichtig‘ sei verstanden in dem Sinne, daß verschiedene Beurteiler jeder Frage gleiche Bedeutsamkeit für den Gesamtzusammenhang zuerkennen.)

Das *Grundproblem* dürfte darin liegen, daß - bei komplexen Gegebenheiten - der ‚Sachverhalt‘ von sich aus *unabschätzbar viele Freiheitsgrade einer Ge-*

staltung zuläßt und somit eine Aufteilung in unterschiedliche Einheiten verträgt.

Es kehrt das Grundproblem jeder Itemgenerierung wieder das darin besteht, Items streng theoretisch abzuleiten und sie regelhaft zu vervielfältigen.

5.3 Analyse kriteriumsorientierter Testaufgaben

Damit die Aufgaben das Kriterium angemessen repräsentieren und einen zu treffenden individuellen Score ergeben, müssen Items und Test valide, reliabel und objektiv sein. Bezug genommen wird also auf die klassischen Testgütekriterien.

„Was von der klassischen Testtheorie in eine kriteriumsorientierte Testtheorie übernommen werden kann, das sind die wichtigsten Testgütekriterien wie Gültigkeit, Objektivität und Zuverlässigkeit, d. h. wir verlangen erstens, daß ein kriteriumsorientierter Test auch das mißt, was er zu messen vorgibt, daß zweitens verschiedene Beurteiler bei Einsatz des gleichen kriteriumsorientierten Tests zu gleichen Ergebnissen kommen und daß drittens die erhaltenen Testwerte nur mit einem geringen Meßfehler behaftet sind“ (Fricke, 1974, 19).

Alternative Algorithmen: Bei Bestimmung von Validität, Reliabilität und Objektivität kann bei kriteriumsorientierten Tests der Fall eintreten, daß zwischen den Test-Scores der Probanden keine Varianz mehr auftritt - dann nämlich, wenn alle Probanden das Kriterium vollständig erreichen. Dieser Fall ist in der klassischen Testtheorie nicht vorgesehen.

Denn die klassische Testtheorie beruht auf einem differentiellen Ansatz: Ein ‚klassisches‘ Instrument ist dazu bestimmt, *Unterschiede zwischen Probanden* zu ermitteln. Wo keine Unterschiede auftreten, sind ihre Algorithmen nicht definiert.

Darum wurden für kriteriumsorientierte Leistungsmessung alternative Algorithmen entwickelt, die es erlauben, auch dann Gütekriterien zu bestimmen, wenn zwischen den Scores der Probanden keine Varianz mehr auftritt. Klauer führt verschiedene Beispiele an (1987, 51-57).

Hier sei **ein** Beispiel erwähnt: der **Ü-Koeffizient** von Fricke (1974, 40-43), der auch dann ‚funktioniert‘, wenn sich zwischen den Probanden keine Varianz mehr zeigt. Wie der Ü-Koeffizient zu verstehen und zu berechnen ist, wird ein Beispiel bei dem Thema ‚Objektivität kriteriumsorientierter Tests‘ verdeutlichen (S. 138).

Besprochen seien die drei klassischen Gütekriterien in ihrem Bezug zur kriteriumsorientierten Messung. An der Spitze stehe der Kennwert, der die kriteriumsorientierte Messung charakterisiert:

- Validität (5.3.1),
- Reliabilität (5.3.2),
- Objektivität (5.3.3).

5.3.1 Validität

Ein kriteriumsorientierter Test definiert sich von seinem Inhalt her. Valide ist er dann, wenn seine Items die Inhalte des Kriteriums vollständig enthalten (enumerative Aufzählung) oder repräsentativ abbilden (Angabe mengenbildender Merkmale).

Kasten 5-5:

Überprüfung der Kontentvalidität bei kriteriumsorientierten Tests - Beispiele (Klauer, 1987, 47-50)

Zwei Fragen werden gestellt:

1. Gehören generierte Items zur definierten Grundmenge?
2. Stimmen die tatsächlichen Itemquoten mit den vorgesehenen Quoten überein?

Zur 1. Frage: Gehören generierte Items zur definierten Grundmenge?

Es wird eine Liste der Merkmale angelegt, die ein Item aufweisen muß. Experten vergleichen jedes Item mit dieser Merkmalsliste.

- Die Entscheidung „**Merkmale vorhanden**“ sei kodiert mit 1,
- die Entscheidung „**Merkmale nicht vorhanden**“ mit 0.

Der Grad der Experten-Übereinstimmung sei beispielsweise mit dem Ü-Koeffizienten ausgedrückt.

Items	Merkmale			
	1	2	3	4
1	1	1	0	1
2	0	1	1	0
		usw.		

Zur 2. Frage: Stimmen die tatsächlichen Itemquoten mit den vorgesehenen Quoten überein?

Für Teilmengen A, B, C von Items werden Quoten vorgesehen. Experten ordnen generierte Items den Teilmengen zu. Die Übereinstimmung kann wieder mit einem Koeffizienten, beispielsweise dem U-Koeffizienten, geprüft werden.

Darüber hinaus kann mit einem „Test auf Güte der Anpassung“, z.B. mit dem χ^2 -Test, geprüft werden, wie weit die tatsächlichen Quoten den vorgesehenen Quoten entsprechen.

Von den Validitätsarten hat daher die **Inhaltsvalidität** Vorrang; sie „ist eine Eigenschaft von Tests als Aufgabenmengen und zunächst unabhängig von jeder Empirie“ (Klauer, 1987, 13). „Kontentvalidität ist das unterscheidende Merkmal, das kriteriumsorientierte von anderen Tests abhebt“ (Klauer, 1987, 8). Alle Methoden, welche eine Kontentvalidierung ermöglichen, sind anwendbar

(S. 95): theoriegeleitete Präzisierung der benötigten Konstrukte, Expertenbefragung usw.

Im Dienst der inhaltlichen Validierung stehen die Erzeugung von Items mittels generativer Regeln und die Überprüfung, ob die hergestellten Items zum definierten Kriterium gehören und in den vorgesehenen Quoten vertreten sind.

Den Schritt der Item-Erzeugung haben Kasten 5-1 bis 5-4 veranschaulicht, den der Item-Überprüfung soll Kasten 5-5 skizzieren.

Experimentelle Überprüfung der Kontentvalidität

Inhaltsvalidität läßt sich auch „experimentell klären, indem zwei unabhängig voneinander arbeitenden Testentwicklungsgruppen erstens eine Beschreibung des zugrundeliegenden Universums und *zweitens* ein Satz von Regeln, wie Testaufgaben zu generieren sind, gegeben werden. Beide Tests werden den gleichen Personen vorgelegt und die resultierenden Testwerte werden miteinander verglichen. Bei einer hohen Übereinstimmung wurde man auf hohe Inhaltsvalidität dieses Tests schließen“ (Fricke, 1974, 45).

Andere Aspekte inhaltlicher Validität

Weitere Aspekte der Validierung kommen in den Blick, wenn gefragt wird

- nach der „**pädagogischen Bedeutsamkeit**“ eines Tests (eine Frage, die in „herkömmlichen Validitätsstudien“ nicht vorkommt),
- nach **Konstrukten**, die einem Verfahren zugrundeliegen,
- nach der **Entscheidungsrelevanz** für eine Zuordnung zu verschiedenen (pädagogischen oder therapeutischen) Behandlungen (Fricke, 1974, 46).

Neben der Kontentvalidität spielen **kriterien- und konstruktbezogene Validität eine geringere Rolle** (Klauer, 1987, 43-46). Der Grund liegt darin, daß die Item-Generierung von der Kontentvalidität her konzipiert ist und ‚eigentlich‘ nur zu überprüfen ist, ob die tatsächlichen Items diesem Konzept entsprechen (wie es z. B. die Schritte in Kasten 5-5 veranschaulichen). Nicht zu überprüfen ist ‚eigentlich‘, ob die Items anderen Kriterien entsprechen, wie es kriteriums- und konstruktbezogene Validierung vorsehen.

Dennoch werden auch kriteriumsbezogene Validität und Konstruktvalidität ermittelt, ebenso die Gütekriterien Reliabilität und Objektivität (Klauer, 1987, 39-46):

„Die empirische Überprüfung kontentvalider Aufgabenstichproben“ wird empfohlen, „weil sich mit ihrer Hilfe ‚schlechte‘ Aufgaben entdecken lassen“. Der Grund liegt darin, „daß ein subjektives Moment trotz strenger

Erzeugungs- und Samplingvorschriften bei der Erarbeitung der Aufgaben bleibt. Aus diesem Grunde muß man mit Fehlern aller Art bei der Anwendung der Vorschriften rechnen, und die empirische Überprüfung der Testaufgaben kann eine Möglichkeit sein, solchen Fehlern auf die Spur zu kommen" (Klauer 1987, 40).

5.3.2 Reliabilität

Ein kriteriumsorientierter Test soll reliabel, d.h. möglichst fehlerfrei sein. Ein Maß wie der \ddot{U} -Koeffizient erlaubt es, die Reliabilität nach den vier klassischen Paradigmen zu bestimmen: als Retest-, Paralleltest- und Halbierungs-Reliabilität sowie als interne Konsistenz.

Doch ist zu beachten: „Was man unter der Reliabilität eines kriteriumsorientierten Tests sinnvollerweise zu verstehen habe, ist noch immer nicht ganz ausdiskutiert“ (Klauer 1987, 84).

5.3.3 Objektivität

Als Objektivität sei hier nur die Übereinstimmung der Auswerterurteile erwähnt. Bei denselben Probanden, die dasselbe Verfahren bearbeiten, sollen die *Auswerterurteile* identisch ausfallen. Der Grad der Übereinstimmung sei mit dem \ddot{U} -Koeffizienten überprüft (Fricke, 1974, 42-43). - Er läßt eine Berechnung auch dann noch zu, wenn in den Test-Scores zwischen den Probanden keine Varianz mehr auftritt, etwa weil alle Probanden das Kriterium vollständig erreicht haben.

Der \ddot{U} -Koeffizient ist definiert in Gleichung (I):

$$(I) \quad \ddot{U} = 1 - \frac{QS_{inh}}{QS_{max}}$$

Es bedeuten:

\ddot{U} : Übereinstimmungskoeffizient,
 QS_{inh} : Quadratsumme ‚innerhalb‘,
 QS_{max} : Maximal mögliche Quadratsumme.

Kasten 5-6 soll die Berechnung des \ddot{U} -Koeffizienten an einem Zahlenbeispiel veranschaulichen.

Kasten 5-6:
Höhe der Auswerterobjektivität, ausgedrückt durch den Ü-Koeffizienten

Pbn :	Probanden				
N :	Zahl der Pbn,		hier: N = 7		
K :	Zahl der Auswerter,		hier: K=4 (A bis D),		
P :	Summenscore je Zeile,		hier: 6, 14...7, 4		
Berechnungsschritte: Unterhalb des Kastens!					
Auswerter					
Pbn	A	B	C	D	P
1	1	2	1	2	6
2	3	4	4	3	14
3	2	2	2	1	7
4	4	3	4	3	14
5	3	5	3	5	16
6	2	1	2	2	7
7	1	1	1	1	4

Berechnungsschritte

Es folgen Berechnungsschritte zur Bestimmung der ‚Quadratsumme innerhalb‘ (QS_{inh}) und der ‚maximal möglichen Quadratsumme‘ (QS_{max}) aus der Formel (I).

$$(A) \quad QS_{inh} = \sum X^2 - \frac{\sum P^2}{K}$$

$$(B) \quad QS_{max} = N \cdot K \cdot (X^2_{max}/4)$$

Es bedeuten:

X : Einzelner Meßwert (hier: X läuft von 1 bis 5),

X_{max} : Höchster Itemwert (hier: 5),

N : Zahl der Probanden (hier: 7),

K : Zahl der Auswerter (hier: 4),

P : Summenscore je (beurteiltes Objekt, hier: je) Proband.

Die Werte aus Kasten 5-6, eingesetzt in (A) und (B), ergeben:

$$(A) \quad QS_{inh} = (1^2 + 2^2 + 1^2 + 2^2 \dots + 1^2 + 1^2 + 1^2 + 1^2) - \frac{6^2 + 14^2 \dots + 7^2 + 4^2}{4}$$

$$= 208 - 798/4 = 8.5$$

$$(B) \quad QS_{max} = 7 \cdot 4 \cdot (5^2/4) = 175$$

Einsetzen von (A) und (B) in Gleichung (I) ergibt:

$$\ddot{U} = 1 - \frac{QS_{inh}}{QS_{max}} = 1 - \frac{8.5}{175} = 0.95$$

Ergebnis: Der Koeffizient $\ddot{U} = 0.95$ zeigt eine hohe Übereinstimmung zwischen den Auswertern an.

Für **dichotome** Daten gibt Fricke (1974, 41) eine **Prüfgröße** an, nicht für mehrfach abgestufte Daten (wie in unserem Beispiel). Lindner (1980) berichtet über eine schärfere Prüfgröße.

Frickes Prüfgröße für dichotome Daten lautet:

$$X^2 = \frac{4N}{K(N-1)} (K \cdot \Sigma X - \Sigma X^2) \quad \text{bei df} = K(N-1)$$

HINWEIS: Dem \ddot{U} -Koeffizienten könnte in der klassischen Testtheorie eine Formel entsprechen, welche der Konsistenzschätzung dient: $rk = 1 - (s_{\text{inh}}^2/s_{\text{zw}}^2)$. Es bedeuten: s_{inh}^2 : 'Varianz innerhalb', s_{zw}^2 : 'Varianz zwischen' (siehe S. 87).

- **Problem:** Wenn zwischen den Probanden keine Varianz auftritt, wenn also s_{zw}^2 den Wert Null annimmt, ist die Formel nicht mehr definiert.

5.4 Schluß vom Testscore auf die Fähigkeit eines Probanden

Wie soll festgelegt werden, in welchem Grade ein Lehr- oder ein Therapieziel, also das Kriterium, erreicht ist?

Bei normorientierten Tests wird der Test-Score eines Probanden verglichen mit den Werten einer Eichstichprobe. Dann wird beispielsweise entschieden, ob der Score für eine durchschnittliche, über- oder unterdurchschnittliche Fähigkeit des Probanden spricht. - Kapitel 4.4 beschreibt diese Prozedur (S. 111).

Bei **kriteriumsorientierten Tests** wird der individuelle Score verglichen mit dem Kriterium: Die Nähe zum Kriterium (etwa einem Lehr- oder einem Therapieziel) gibt an, wie eine Leistung zu klassifizieren ist, wie hoch demnach die Fähigkeit eines Probanden einzustufen ist.

Es sind unterschiedliche Vorschläge gemacht worden, die eine Entscheidung über diese 'Nähe' erlauben. Dazu drei Beispiele:

- Einstufige Entscheidung:

Festlegung eines kritischen Punktwertes (5.4.1),

- Mehrstufige Entscheidung:

Festlegung von Entscheidungsintervallen (5.4.3),

- Entscheidungen mit Bestimmung eines Vertrauensbereiches (5.4.2).

Wir veranschaulichen die drei Beispiele an dem Paradigma eines Tests, der 40 Items enthält.

5.4.1 Einstufige Entscheidung:

Festlegung eines kritischen Punktwertes

Ein ‚kritischer Punktwert‘ (Cut-Off-Point) wird festgelegt. Erreicht ein Proband den Wert, so gilt das Urteil: ‚Kriterium erreicht‘. Bleibt ein Proband unterhalb des Wertes, so bedeutet dies: ‚Kriterium nicht erreicht‘.

Beispiel: *Bei einem kriteriumsorientierten Test, der aus 40 Items besteht, werde festgelegt: Wer zwei Drittel der Aufgaben, also 27 Items, löst, hat das Ziel erreicht.*

Zentrales Problem: Wie kann der Testautor den kritischen Punktwert rechtfertigen? Um das Beispiel weiterzuführen: Wie kann der Autor rechtfertigen, daß zwei Drittel der Lösungen genügen, um das Kriterium als erreicht zu deklarieren?

Es wird ähnlich verfahren wie bei Schulzeugnissen. Als kritischer Punktwert wird in der Regel die Note „Ausreichend“ (vier) festgelegt. Noten, die gleich hoch oder höher ausfallen, besagen: ‚Ziel erreicht‘. Noten, die unter „Ausreichend“ liegen, signalisieren: ‚Ziel verfehlt‘. -Für Noten erhebt sich die Frage: Wie läßt sich der ‚Schnitt‘ bei der Note für begründen?

5.4.2 Mehrstufige Entscheidung:

Festlegung von Entscheidungsintervallen

Statt festzulegen, daß ein einzelner Punktwert darüber entscheidet, ob ein Kriterium erreicht ist, läßt sich der Vorgang auch abstufen. Der Testautor unterteilt den Abstand zwischen erreichtem und kritischem Wert in mehrere Intervalle (Fricke, 1974, 97).

Beispiel: *Die Entscheidungsstufen beziehen sich wieder auf einen Test mit 40 Aufgaben. Für unterschiedliche Anteile ‚gelöster Aufgaben‘ seien unterschiedliche Noten vergeben.*

Festgelegt wird etwa:

- *Eine Mindestleistung ist erbracht, wenn mehr als 66, aber weniger als 75 Prozent der Aufgaben gelöst sind; dafür stehe die Note „Ausreichend“.*
- *Die höchste Leistung ist erbracht, wenn mehr als 95 Prozent der Aufgaben gelöst werden; dafür stehe die Note „Sehr gut“.*

Auf diese Weise kann ein mehrstufiges Punktesystem erstellt werden, das die Nähe zum Kriterium in mehreren Intervallen angibt.

Kasten 5-7 veranschaulicht das Beispiel.

Kasten 5-7:
Mehrstufige Entscheidung, veranschaulicht an Notenstufen

40 Aufgaben eines Tests sind vorgegeben. Es wird festgelegt, wieviele Aufgaben für welche Note zu lösen sind.		
<i>Prozent gelöster Aufgaben</i>	<i>Anzahl gelöster Aufgaben</i>	<i>Note</i>
Unter 66	Unter 27	Nicht ausreichend
66 bis 75	27 bis 30	Ausreichend
76 bis 82	31 bis 33	Befriedigend
83 bis 94	33 bis 37	Gut
95 bis 100	38 bis 40	Sehr gut

Problem: Kann der Testautor überzeugend rechtfertigen, warum er bestimmte Lösungsanteile bestimmten Intervallen zuordnet? Kann er nachvollziehbar begründen, daß ein Proband 95 Prozent der Aufgaben lösen muß, um die Note „Sehr gut“ zu erreichen?

5.4.3 Entscheidungen mit Vertrauensbereich

Statt kritische Einzelpunkte oder kritische Intervalle festzulegen, kann *der* Testautor auch einen Bereich angeben, zu dem der ‚Fähigkeitswert‘ eines Probanden gehören muß - er **kann ein Vertrauensintervall bestimmen, das den Kriteriumswert einschließt**.

Wir gliedern dieses Teilkapitel in zwei Abschnitte:

- Bedeutung eines Vertrauensbereiches (5.4.3.1),
- Berechnung eines Vertrauensbereiches (5.4.3.2).

Die beiden Fragen überschneiden sich, darum sind inhaltliche Wiederholungen zu erwarten.

5.4.3.1 Bedeutung eines Vertrauensbereiches

Die Berechnung eines Vertrauensbereiches hat die Funktion, aufmerksam zu machen auf die punktuelle Unschärfe eines einzelnen Test-Scores. Zentrales Ziel ist es, anzuzeigen: Ein ermittelter empirischer Test-Score umschreibt *nur mit einer gewissen Wahrscheinlichkeit* den Bereich, in dem der *wahre Wert* liegt.

Vom Gegenteil her formuliert: die Berechnung eines Vertrauensbereiches warnt den Anwender davor, seinen aktuell beobachteten Wert als eine ‚sichere Information‘ zu betrachten. Die Berechnung soll ihn dazu anleiten, in seiner Entscheidung mit hohen Graden an Unsicherheit zu rechnen.

Ob die konkrete Berechnung eines Vertrauensbereiches mittels Standardmeßfehlers selber die ‚Wahrheit‘ anzeigt: diese Frage ist ihrerseits theoretisch

*strittig. Der hauptsächliche Einwand artikuliert sich in der Frage, ob sich Statistiken, die an einer **Gruppe** gewonnen wurden, übertragen lassen auf eine Einzelperson.*

Zulässig ist eine Übertragung nur dann, wenn geklärt ist, daß die Einzelperson typisch ist für die Gruppe, an der Statistiken erhoben wurden.

Gegen die Einwände sei aber *auch* festgehalten: Was die Berechnung eines Vertrauensbereiches rechtfertigt, ist ihr Charakter als Appell, der ausdrücken soll: *Ein aktueller Testwert*

- ist *kein eindeutiger Indikator* der Ausprägung eines Persönlichkeitsmerkmals,
- sondern läßt sich nur betrachten als *ein „Repräsentant“*, der anzeigt, welchem *Umfeld* ein Merkmal zuzuordnen ist.

Der Vertrauensbereich soll dieses *Umfeld* abgrenzen, er umreißt das Intervall, in dem *mit einer statistischen Wahrscheinlichkeit* der sogenannte ‚wahre Wert‘ zu erwarten ist. ‚Nach oben‘ und ‚nach unten‘ stecken Grenzen den Umfang des Vertrauensintervalles ab. Die Angabe, daß der wahre Wert innerhalb des Vertrauensbereiches liegt, signalisiert eine statistische *Wahrscheinlichkeit für* eine Entscheidung - mehr nicht; sie warnt vor allzu hoher Sicherheit.

5.4.3.2 Berechnung eines Vertrauensbereiches

Das Modell, auf dem die Berechnung eines Vertrauensbereiches beruht, wurde dargestellt in Kapitel 4.3.2.5 (S. 89). *Der Grundgedanke sei kurz wiederholt:* Ein Proband löse 30 Items in einem Test, der 40 Items umfaßt. Sein Test-Score ist jedoch fehlerbehaftet. Darum stellt sich die Frage: *Wie weit erstreckt sich der Fehlerbereich ,über 30 hinaus nach oben‘ und ,unter 30 nach unten‘?*

Dieser Suchbereich heißt der Vertrauensbereich (VB), darstellbar in der Formel:

$$VB = X \pm z_a \cdot \text{Fehler}$$

Es bedeuten:

- x : erreichter Test-Score,
 z_a : gewählte Wahrscheinlichkeit, mit welcher der Fehlerbereich ‚abgesteckt‘ wird, also der z-Wert für den a-Fehler (z.B. $p \ 55 \% \Rightarrow z = 1.96$).

Der Standardmeßfehler wie ihn die klassische Testtheorie festlegt, eignet sich bei kriteriumsorientierter Messung nicht zur Schätzung des Vertrauensbereiches - aus dem immer wieder genannten Grund: Es kann vorkommen, daß keine Varianz ‚zwischen den Probanden‘ auftritt, ein Fall, für den klassische Algorithmen nicht ausgelegt sind.

Darum wurden andere Modelle entworfen, die es erlauben, Vertrauensbereiche zu berechnen. Wir stellen nur *ein* Beispiel vor: die Schätzung des Vertrauensintervalles nach dem Binomialmodell.

Fehlerschätzung nach dem Binomialmodell

Klauer schildert die Voraussetzungen und die Möglichkeiten einer Anwendung und Erweiterung des Binomialmodells sehr differenziert, er veranschaulicht sie an vielfältigen Beispielen (1987, 137-250). *Aus Klauers ausführlicher Darstellung referieren wir nur einen Gedankengang.*

Definition: Die Binomialformel lautet bekanntlich $(p + q)^n$. Die Elemente p und q beziehen sich auf die Wahrscheinlichkeit, mit der komplementäre Ereignisse eintreten. Trete ein Ereignis mit der Wahrscheinlichkeit $p = 0.75$ ein, dann gilt für das Eintreten des komplementären Ereignisses q : $q = 1 - 0.75$.

Bei einer Testkonstruktion bezeichne beispielsweise

- p die Wahrscheinlichkeit des Ereignisses: ‚Item gelöst‘,
- q dagegen die Wahrscheinlichkeit des Ereignisses: ‚Item nicht gelöst‘.

Der Exponent ‚ n ‘ benennt die Anzahl der ‚Versuche‘, in denen das Ereignis p auftritt.

Kasten 5-8:

Binomialmodell - Ermittlung der Wahrscheinlichkeit eines Ereignisses - Grundformel

Wir schildern die Wahrscheinlichkeit, ob ein bestimmtes Ereignis eintritt, am Beispiel eines Tests. Die allgemeine Schätzformel lautet:

$$p(x | n) = \frac{n!}{x! \cdot (n-x)!} \pi^x \cdot (1-\pi)^{n-x}$$

Es bedeuten:

- n : Zahl der ‚Versuche‘, hier: Gesamtzahl der Items,
- x : Zahl der Ereignisse, die mit der Wahrscheinlichkeit p eintreten, hier: Anzahl gelöster Items,
- π : Wahrscheinlichkeit, mit der ein Ereignis eintritt, hier: Wahrscheinlichkeit, mit der ein Item gelöst wird (geschätzt über den Schwierigkeitsindex p),
- $1-\pi$: Wahrscheinlichkeit des komplementären Ereignisses.

Beispiel: Ein Test habe 10 Items. Jedes Item habe die Wahrscheinlichkeit $p = 0.50$, gelöst zu werden. Diese Wahrscheinlichkeit werde abgeleitet aus dem Schwierigkeitsindex $p = 0.50$. Demnach gilt: $p = 0.5$ und $q = (1-p) = 0.5$. Wie groß ist die Wahrscheinlichkeit, 5 Items zu lösen? - Es gilt: $n = 10$; $x = 5$; $p = 0.5$;

Einsetzen:

$$p(5 | 10) = \frac{10!}{5! (10-5)!} \cdot 0.5^5 \cdot (1-0.5)^{10-5}$$

$$p(5 | 10) = 252 \cdot 0.031 \cdot 0.031 = 0.2461$$

Die Wahrscheinlichkeit, fünf der zehn Testaufgaben zu lösen, beträgt $p \approx 0.25$.

Beispiel: In einer Urne liegen 5 schwarze und 5 weiße Kugeln. ,n‘ bezeichne 10 Versuche, in denen eine schwarze oder eine weiße Kugel entnommen wird,

Wie die Wahrscheinlichkeit eines Ereignisses nach dem Binomialmodell ermittelt wird, daran erinnert Kasten 5-8.

Voraussetzungen: Bei Verwendung des Binomialmodells gelten drei Voraussetzungen (Klauer. 1987. 138):

1. Komplementarität der Ereignisse: Die Ereignisse, um die es geht, stehen zueinander in einem komplementären Verhältnis; es kann - bei *einem* Versuch - nur *eines* der beiden Ereignisse eintreten.

Beispiel: In einer Urne liegen nur schwarze und weiße Kugeln; bei einem Vorgang wird jeweils nur *eine* Kugel entnommen: eine schwarze oder eine weiße.

2. Lokale stochastische Unabhängigkeit der Ereignisse: Das Eintreten des *einen* Ereignisses ,beeinflußt‘ nicht das Eintreten des *anderen* Ereignisses.

Beispiel: Wer der Urne beim ersten Male eine weiße Kugel entnimmt, entscheidet damit nicht, welche Kugel er beim zweiten Male zieht,

3. Gleichwahrscheinlichkeit der Ereignisse: Welches der beiden komplementären Ereignisse eintritt, ist bei jedem Vorgang gleich wahrscheinlich.

Beispiel: Nach jeder Entnahme wird die entnommene Kugel in die Urne zurückgelegt; bei erneuter Entnahme ist somit für alle Kugeln das Ereignis ,weiße Kugel‘ gleich wahrscheinlich wie das Ereignis ,schwarze Kugel‘.

Ermittlung des Vertrauensbereiches bei kriteriumsorientierten Tests

Der Vertrauensbereich wird auf zweifache Weise bestimmt:

- durch Ablesen in einer Tabelle (A),
- durch Berechnung mittels Näherungsformeln (B).

(A) Ablesen in einer Tabelle

Vertrauensintervalle sind in Tabellen festgehalten, so bei Klauer (1987, 284-291) oder bei Fricke (1974, 105-108). Kasten 5-9 bringt einen Auszug.

Kasten 5-9:
Vertrauensgrenzen für n = 40 Ereignisse,
hier: für 40 Items / Auszug aus Klauer (1987, 288)

Erläuterungen: Es geht um einen ‚Test‘, der 40 Items umfaßt.

- **Spalte 1** gibt die Zahl der gelösten Items an.
- **Spalte 2** gibt an, wie hoch der Prozentanteil ‚der gelösten Items‘ an der Gesamtzahl der 40 Items ausmacht.
- **Spalte 3** gibt an, wo für die Zahl der gelösten Items die untere und die obere Vertrauensgrenze liegt - mit $p \leq 5 \%$.

<i>Spalte 1</i>	<i>Spalte 2</i>	<i>Spalte 3</i>
<i>Zahl gelöster Items</i>	<i>Prozent gelöster Items</i>	<i>Vertrauensbereich: Grenzen bei $p \leq 5 \%$</i>
0	0.00	0.00 - 8.81
1	2.50	0.06 - 13.16
2	5.00	0.61 - 16.92
3	7.50	1.57 - 20.39
4	10.00	2.79 - 23.66
5	12.50	4.19 - 26.80
...		
30	75.00	58.80 - 87.31
31	77.50	61.55 - 89.16
32	80.00	64.35 - 90.95
33	82.50	67.22 - 92.66
34	85.00	70.16 - 94.29
35	87.50	73.20 - 95.81
36	90.00	76.34 - 97.21
37	92.50	79.61 - 98.43
38	95.00	83.08 - 99.39
39	97.50	86.84 - 99.94
40	100.00	91.19 - 100.00

Zwei Beispiele zu Kasten 5-9:

1. Frau Müller habe 30 von 40 Items gelöst. *Nachlesen in Kasten 5-9 ergibt:*
 - **Spalte 1 und Spalte 2:** 30 gelöste Items repräsentieren 75 Prozent der 40 angebotenen Items.
 - **Spalte 3:** Die Vertrauensgrenzen liegen mit $p \leq 5 \%$
 - ⇒ ‚nach unten‘ bei 58.80 Prozent,
 - ⇒ ‚nach oben‘ bei 87.31 Prozent.

Demnach gilt: Der wahre Wert von Frau Müller mit 30 Lösungen liegt zwischen rund 59 und rund 87 Prozent der Lösungen (mit $p \leq .5 \%$).
2. Herr Allers habe 31 von 40 Items gelöst, Frau Bach dagegen 32 Items. Wie groß ist die Wahrscheinlichkeit mit $p \leq 5 \%$, daß die beiden zu jenen Probanden gehören, die ‚in der Regel‘ 90 Prozent der Aufgaben lösen?
 - Wer 31 Items von 40 löst, liegt nach Kasten 5-9 in einem Intervall, das 61 bis 89 Prozent der Lösungen umfaßt.
 - Wer dagegen 32 Items von 40 löst, liegt in einem Intervall, das von 64 bis 91 Prozent reicht.

Demnach gilt: Herr Allers mit 31 Lösungen liegt außerhalb des angesetzten Vertrauensbereiches (mit $p = 5\%$). Dagegen liegt Frau Bach mit 32 Lösungen innerhalb des angesetzten Vertrauensbereiches (mit $p \leq 5\%$).

(B) Berechnung eines Vertrauensbereiches

Wir wiederholen die allgemeine Formel, nach welcher sich der Vertrauensbereich (VB) bestimmt:

$$VB = X \pm z_{\alpha} \cdot \text{Fehler}$$

Wir fügen die Näherungsformel an, die Klauer anführt (1987, 150):

$$VB = (A - B) \pm \sqrt{B \cdot [2 - (A - B) - A]}$$

A und B bestimmen sich wie folgt:

$$A = \frac{X_i + 0.5 \pm 0.5 + \frac{z^2}{4}}{n + 1}$$

$$B = \frac{z^2}{2} \cdot \frac{X_i + 0.5 \pm 0.5}{(n + 1)^2}$$

In A und B bedeuten:

X_i : Zahl der gelösten Items, z.B. 30,

n : Zahl aller Testitems, z.B. 40,

z : z-Wert für α -Fehler, z.B. für $p \leq 5\% \Rightarrow z = 1.96$.

Beispiel zur Berechnung eines Vertrauensbereiches

Aufgabe: Für das Beispiel 1 (Frau Müller) „Vertrauensintervalle ablesen in Tabellen“ kennen wir die obere und untere Vertrauensgrenze, nämlich 58.80 % (unten) und 87.31 % (oben) - abgelesen in Kasten 5-9.

Jetzt ermitteln wir die untere und die obere Vertrauensgrenze mit der Näherungsformel.

Frau Müller hat 30 von 40 Items gelöst. Als Restwahrscheinlichkeit sei das 5-Prozent-Niveau gewählt. Somit gilt: $n = 40$; $X_i = 30$; $z_{\alpha} = 1.96$.

Berechnung: Wir berechnen zuerst die untere, dann die obere Grenze.

1. Untere Grenze p_u :

Zunächst werden A_u und B_u bestimmt.

$$A_u = \frac{30 + 0.5 - 0.5 + \frac{1.96^2}{4}}{40 + 1} = 0.755$$

$$B_u = \frac{1.96^2}{2} \cdot \frac{30 + 0.5 - 0.5}{(40 + 1)^2} = 0.034$$

Die Resultate von A_u und B_u werden eingesetzt in die Gesamtformel für die untere Grenze p_u :

$$p_u = [0.755 - 0.034] - \sqrt{0.034 \cdot [2 - (0.755 - 0.034) - 0.755]}$$

$$p_u = 0.721 - 0.133 = 0.588 \Rightarrow 58,8 \%$$

Resümee: Die untere Grenze liegt bei $p_u = 0,588$.

Obere Grenze p_o :

Zunächst werden A_o und B_o bestimmt:

$$A_o = \frac{30 + 0.5 + 0.5 + \frac{1.96^2}{4}}{40 + 1} = 0.779$$

$$B_o = \frac{1.96^2}{2} \cdot \frac{30 + 0.5 + 0.5}{(40 + 1)^2} = 0.035$$

Die Werte von A_o und B_o werden eingesetzt in die Gesamtformel für die obere Grenze p_o :

$$p_o = [0.779 - 0.035] = \sqrt{0.035 \cdot [2 - (0.779 - 0.035) - 0.779]}$$

$$p_o = 0.744 + 0.129 = 0.873 \Rightarrow 87,3 \%$$

Resümee: Die obere Grenze liegt bei $p_o = 0.873$.

Vergleich von Tabellen- und Rechenergebnis: In Beispiel I (Frau Müller: 30 Items von 40 Items gelöst) stimmen die berechneten Werte $p_u = 58,8 \%$ und $p_o = 87,3 \%$ mit den Tabellen- Werten überein.

5.5 Beitrag zu Diagnostik und Intervention

Der kriteriumsorientierte Ansatz ist eigens zum Zweck der Interventionskontrolle konzipiert worden: vor allem in Pädagogischer und Klinischer Psychologie. Die diagnostischen und die interventiven Schritte gehen darum ineinander über. Dafür zwei Beispiele:

- Die kriteriumsorientierte Messung ermöglicht:

- ⇒ eine prägnante *Umschreibung des Zieles*, etwa für eine pädagogische oder für eine therapeutische Intervention,
- ⇒ eine exakte *Angabe der ‚Mittel‘*, die zur Erreichung des Zieles eingesetzt werden sollen.
- Die kriteriumsorientierte Messung ermöglicht:
 - ⇒ die Angabe, *wie nahe* ein Proband dem angestrebten Ziel gekommen ist,
 - ⇒ den *Vergleich mit anderen Probanden*, die dem gleichen Ziel zustreben.

5.6 Zusammenfassung zu Kapitel 5

Die kriteriumsbezogene Leistungsmessung läßt sich betrachten als eine Erweiterung und Ergänzung der klassischen Testtheorie. Individuelle Leistung wird nicht verglichen mit den Leistungen einer Eichstichprobe (realnorm-orientierte Messung), sondern mit einem inhaltlich festgelegten Kriterium (ideálnorm-orientierte Messung).

Kriteriumsorientierte Tests basieren auf einer präzise definierten Grundmenge von Aufgaben“ (Klauer, 1987, 59): In dieser Umschreibung spiegelt sich ihre Eigenart und - auch ihre Problematik (Moser, 1987). Die Problematik ergibt sich aus der Schwierigkeit, das zugrundeliegende Konstrukt und die zugehörige Grundmenge von Items ‚präzise zu definieren‘.

Ein kriteriumsorientierter Test soll ähnlich hohe Werte in den Gütekriterien aufweisen wie ein normorientierter Test: Er soll objektiv, reliabel, und valide sein. Weil aber die Testaufgaben vom Inhalt des Kriteriums her konzipiert und generiert werden, erhält die Kontentvalidierung Vorrang.

Bei kriteriumsorientierter Leistungsmessung kann der Fall eintreten, daß keine Varianz ‚zwischen den Probanden‘ auftritt; dann ‚versagen‘ die Algorithmen der klassischen Testtheorie. Es wurde ein Algorithmus, der Ü-Koeffizient, vorgestellt, der es auch in diesem Fall ermöglicht, Werte zu ermitteln, die denen der klassischen Testtheorie entsprechen.

In welchem Grade ein Proband das Lehr- oder Therapieziel erreicht hat, wird entschieden durch Vergleich seiner individuellen Leistung mit dem Kriterium: Die Nähe zum Kriterium gibt an, wie hoch die Fähigkeit eines Probanden einzustufen ist. Es wurden drei Möglichkeiten besprochen, diese ‚Nähe‘ zu bestimmen:

1. Einstufige Entscheidungen (Festlegung eines kritischen Punktwertes),
2. Mehrstufige Entscheidungen (Festlegung von Entscheidungsintervallen),
3. Entscheidungen mit Bestimmung eines Vertrauensintervalles.

Zwischen norm- und kriteriumsbezogener Leistungsmessung besteht kein essentieller Unterschied. Denn auch Kriterien werden stichprobenbezogen fest-

gelegt. Wenn etwa ein Kriterium für ‚Rechenfähigkeit in der vierten Klasse‘ bestimmt werden soll, dann müssen sich die ‚Inhalte‘ an der ‚Gruppe der Viertkläßler‘ orientieren; das aber heißt, diese Stichprobe übernimmt die Rolle einer normgebenden Gruppe. Doch ist einzuräumen, daß in den beiden Testklassen (normorientiert vs kriteriumsorientiert) die Art der ‚Orientierung an einer Stichprobe‘ erheblich divergieren.

Darüber hinaus ist zu bedenken: Ein kriteriumsorientierter Test läßt sich in einen normorientierten, ein normorientierter in einen kriteriumsorientierten Test umwandeln.

5.7 Kontrollfragen zu Kapitel 5

- Definition kriteriumsorientierter Leistungsmessung.
- Unterschiede zur klassischen Leistungsmessung.
- Beispiele für Itemgenerierung.
- Rolle der Testgütekriterien.
- Schluß vom Test-Score auf die ‚Fähigkeit‘ eines Probanden.
- Beitrag der kriteriumsorientierten Leistungsmessung zu Diagnostik und zur Intervention.

6. Kapitel

Der Grundgedanke des Rasch-Modells

Mit dem Ziel, Instrumente zu konstruieren, die meßtheoretisch höheren Anforderungen genügen als Verfahren der klassischen Testtheorie, wurden andere Konzeptionen entwickelt, zum Beispiel sogenannte ‚latent-trait‘-Modelle. Instrumente, die ihren Annahmen entsprechen, sollen das Meßniveau von Intervall-, vielleicht sogar von Rationalskalen erreichen.

Diese Modelle beruhen auf der Annahme, „daß die zu erfassenden psychischen Merkmale als latente Dimensionen“ interpretierbar sind. Der beobachtbare Testwert dient als Indikator „für die Beschaffenheit des ‚latent-trait‘-Parameters“ (Michel & Conrad, 1982, 27; vgl. Gigerenzer, 1981; Guthke, Böttcher & Sprung, 1990, 146-170; Kubinger, 1995 a; Wright & Stone, 1979).

Als Beispiel sei das Rasch-Modell besprochen. Wir ordnen den Stoff nach folgenden Gesichtspunkten:

- Modellannahmen (6.1),
- Ausgangsgleichung (6.2),
- Schritte einer Rasch-Skalierung (6.3),
- ergänzende Hinweise zur Endmatrix (6.4),
- Charakteristika einer Rasch-Skala (6.5),
- Beitrag zu Diagnostik und Intervention (6.6).

Das Kapitel schließt mit kritischen Anmerkungen (6.7), einer Zusammenfassung (6.8) und der Vorgabe einiger Kontrollfragen (6.9).

6.1 Modellannahmen

Einige Modellannahmen, von denen Rasch ausgeht, lauten:

1. Jede Person läßt sich hinsichtlich ihrer Fähigkeit, ein bestimmtes Test-Item zu lösen, durch einen Meßwert auf einer eindimensionalen Skala charakterisieren, der **Personenparameter** genannt wird. Für ‚Personenparameter‘ stehe zunächst das Kürzel PP.
2. Jedes Item läßt sich hinsichtlich seiner Schwierigkeit durch einen Meßwert auf einer eindimensionalen Skala charakterisieren, der **Itemparameter** genannt wird. Für ‚Itemparameter‘ stehe zunächst das Kürzel IP.

3. Beide Parameter - Personen- und Itemparameter (PP und IP) - lassen sich gemeinsam **auf einer eindimensionalen Skala** abbilden, so daß immer entscheidbar ist, ob der PP größer oder kleiner als der IP ist oder aber ihm gleich.
4. Der Zusammenhang zwischen der Lösung eines Items und den beiden Parametern (PP, IP) ist **probabilistisch**, d.h. in Abhängigkeit von PP und IP läßt sich dem Ereignis „Item wird gelöst“ eine bestimmte Wahrscheinlichkeit zuordnen.

HINWEIS: Für den Zusammenhang zwischen Personen- und Itemparameter geben Guthke, Böttcher und Sprung in ihrem Lehrbuch eine ‚psychologische‘, vielleicht sogar ‚anthropologische‘ Erklärung (1990, 146):

„Menschliches Verhalten in einer bestimmten Situation ist von soviel zufälligen Faktoren abhängig, daher ‚stochastisch‘ bzw. probabilistisch, daß nach Annahme der Vertreter der PTT“ (probabilistischen Testtheorie) „auch bei genauer Kenntnis des Ausprägungsgrades der latenten Eigenschaft (Indikand) nur Aussagen über die Auftretenswahrscheinlichkeit von manifestem beobachtbarem Verhalten (also Reaktionsverhalten = Testverhalten) gemacht werden können und daher niemals das Testverhalten mit völliger Sicherheit vorhergesagt werden kann.“

Gesucht wird demgemäß eine **Wahrscheinlichkeitsfunktion**, welche es erlaubt, die genannten Annahmen abzubilden:

$$p = f(\text{PP} - \text{IP})$$

In Worten: Die Wahrscheinlichkeit p , daß ein Item gelöst wird, ist eine Funktion (f) des Personen- und des Itemparameters (PP und IP). Je mehr der Personenparameter den Itemparameter ‚übertrifft‘, desto größer ist die Wahrscheinlichkeit, daß die Person das Item löst.

Die gesuchte Funktion soll folgende Eigenschaften haben:

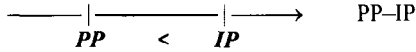
- a) Wenn **IP größer als PP** ist, soll gelten: Je größer die **negative** Differenz $\text{PP} - \text{IP}$ ist, um so kleiner soll die Wahrscheinlichkeit dafür ausfallen, daß die Person j das Item i löst ($p < .50$).
- b) Wenn **PP und IP gleich** sind, soll die Lösungswahrscheinlichkeit fünfzig Prozent betragen ($p = .50$), d.h. in etwa fünfzig Prozent der Fälle soll damit zu rechnen sein, daß die Person j das Item i löst, in etwa fünfzig Prozent der Fälle dagegen nicht.
- c) Wenn **PP größer als IP** ist, soll gelten: Je größer die **positive** Differenz $\text{PP} - \text{IP}$ ist, um so größer soll die Wahrscheinlichkeit dafür sein, daß die Person j das Item i löst ($p > .50$).

Kasten 6-1 soll diese Zuordnung veranschaulichen (vgl. Wright & Stone, 1979, 13).

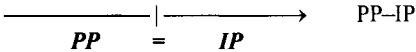
Kasten 6-1:

Wir unterscheiden drei Fälle:

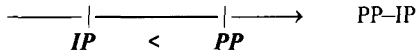
1. Fall: Die Wahrscheinlichkeit, daß die Person j das Item i löst, ist **kleiner als 0.50**.



2. Fall: Die Wahrscheinlichkeit, daß die Person j das Item i löst, ist **gleich 0.50**.



3. Fall: Die Wahrscheinlichkeit, daß die Person j das Item i löst, ist **größer als 0.50**.



Als Kurve stellt sich die gesuchte Wahrscheinlichkeitsfunktion dar, wie in Kasten 6-2 gezeigt. Der Name lautet „**Item-Charakteristik-Kurve**“ (ICC).

Kasten 6-2:

Die gesuchte Wahrscheinlichkeitskurve, genannt „**Item-Charakteristik-Kurve**“ (ICC).

p : Wahrscheinlichkeit \Rightarrow Ordinaten: a, b, c

PP : Personenparameter

IP : Itemparameter

Wie in Kasten 6-1 unterscheiden wir drei Fälle:

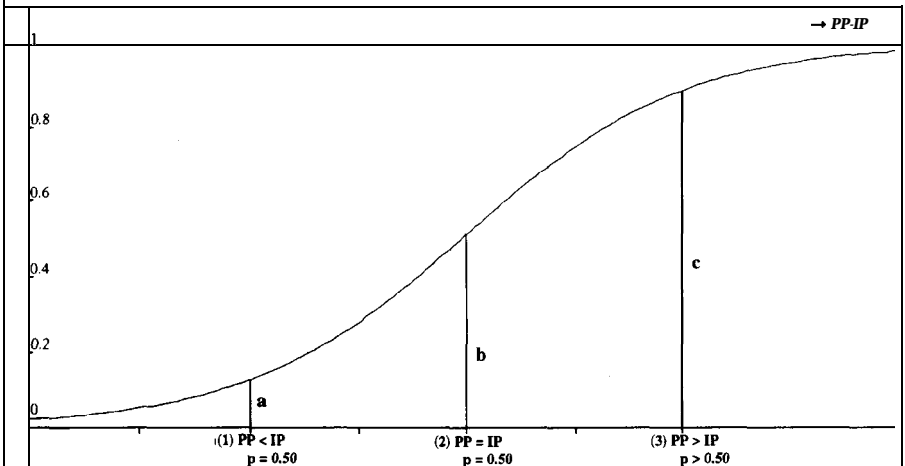
1. $PP < IP: \Rightarrow a \Rightarrow p < 0.50$

2. $PP = IP: \Rightarrow b \Rightarrow p = 0.50$

3. $PP > IP: \Rightarrow c \Rightarrow p > 0.50$

Je mehr der Personenparameter PP den Itemparameter IP übertrifft, desto höher wachsen die Ordinaten: $a < b < c$, desto höher steigt die Wahrscheinlichkeit einer Itemlösung:

$$P_a < P_b < P_c$$



6.2 Ausgangsgleichung

Rasch wählte eine Funktion, die der Normal-Ogive ähnelt, mathematisch jedoch einfacher zu handhaben ist: die **Logistische Funktion** $L(x)$. Sie lautet:

$$L(x) = \frac{e^x}{1 + e^x} = \frac{\exp(x)}{1 + \exp(x)}$$

Welche Eigenschaften hat $L(x)$? Die Logistische Funktion nähert sich dem Wert Null und Eins asymptotisch. Erreichen kann sie weder die Eins noch die Null:

- Den Wert Null kann sie nicht erreichen, weil e^x nie den Wert Null annimmt, gleichgültig, welcher Exponent x gewählt wird. Somit bleibt der gesamte Quotient immer größer als Null.
- Den Wert Eins kann sie nicht erreichen, weil der Zähler immer kleiner bleibt als der Nenner. (Der Nenner ist immer um Eins größer als der Zähler.)

Die Logistische Funktion bildet die Ausgangsgleichung des Rasch-Modells (Rasch, 1960; Michel & Conrad, 1982, 28; Wright & Stone, 1979, 15). Sie liefert die Schätzungen, die angeben, mit welcher Wahrscheinlichkeit ein Item gelöst wird. In die Funktion wird die Differenz PP- IP als Exponent, also für x , eingesetzt.

HINWEIS: Wir schreiben von jetzt an - nach Wright & Stone (1979, 15) - statt PP das Kürzel β (für ability), statt IP das Kürzel δ (für item difficulty) bzw. d und b . Der Exponent ist eine Differenz, wir benennen ihn $x = \beta - \delta$ = dif; somit gilt:

$$p(x = 1|\beta, \delta) = \frac{e^x}{1 + e^x} = \frac{e^{\beta - \delta}}{1 + e^{\beta - \delta}} = \frac{e^{dif}}{1 + e^{dif}}$$

Der Kurvenverlauf sei kurz illustriert:

- Wenn gilt: $\beta < \delta$, etwa $\beta - d = -1$, dann folgt:

$$p = \frac{e^{-1}}{1 + e^{-1}} = \frac{0.3678}{1 + 0.3678} = 0.2689$$

- Wenn gilt: $\beta < \delta$, etwa $\beta - \delta = -0.50$, dann folgt: $p = 0.3775$.
- Wenn gilt: $\beta = \delta$, also $\beta - \delta = 0$, dann folgt: $p = 0.5000$.
- Wenn gilt: $\beta > \delta$, etwa $\beta - \delta = +0.50$, dann folgt: $p = 0.6225$.
- Wenn gilt: $\beta > \delta$, etwa $\beta - \delta = +1$, dann folgt: $p = 0.7311$.

Die Kurve verläuft oberhalb und unterhalb von $p = 0.50$ symmetrisch. Diese Symmetrie soll Kasten 6-3 veranschaulichen.

Kasten 6-3:**Logistische Funktion:** $p = e^x / (1 + e^x)$

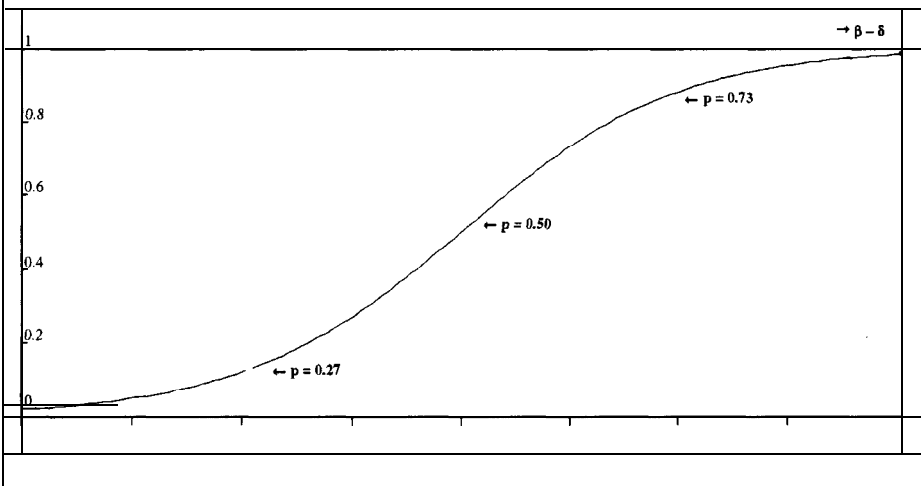
Anwendung auf das Rasch-Modell

 p : Lösungswahrscheinlichkeit $x = \beta - \delta$ β : Personenparameter \rightarrow ability δ : Itemparameter \rightarrow difficultyWir berechnen die Wahrscheinlichkeit p für drei Fälle:

1. $\beta - \delta = -1 \Rightarrow p = 0.27$

2. $\beta - \delta = 0 \Rightarrow p = 0.50$

3. $\beta - \delta = +1 \Rightarrow p = 0.73$

Siehe den laufenden Text!

HINWEIS auf eine andere Schreibweise: Häufig wird die Grundfunktion des Rasch-Modells in anderer Form notiert, beispielsweise: $p = \Theta / (\Theta + \sigma)$. Dabei steht Θ (Theta) für den Personenparameter und σ (Sigma) für den Itemparameter. Diese Formel ist eine Ableitung aus der Logistischen Funktion, sie bezeichnet keine andere Grundgleichung.

Resümee: Werden die Werte der Logistischen Funktion als Wahrscheinlichkeitsangaben interpretiert, ergeben sich die gewünschten Werte zwischen 0 und 1, sofern die Personen- und Itemparameter bekannt sind.

Ansatz einer Rasch-Skalierung

Bei Beginn einer Testkonstruktion sind β und δ unbekannt. Wie kann man die beiden Parameter dann schätzen?

Der Weg sei

- *in den nächsten Abschnitten kurz skizziert,*
- *in den nächsten Teilkapiteln dann ausführlich geschildert.*

Zu Beginn einer Testkonstruktion lassen sich die *Schwierigkeitsindizes* p' der Items berechnen - gemäß dem Ansatz der klassischen Testtheorie, wonach gilt: $p = N_R/N$. (N_R bezeichnet die Zahl der Richtiglöser, N die Zahl der Probanden, die ein Item bearbeitet haben.)

Nun sei **angenommen**, diese „Schwierigkeitsindizes“ p' repräsentierten jene „Wahrscheinlichkeiten“ p' , welche die Logistische Funktion liefern könnte, wenn Personen- und Itemparameter bekannt wären.

Unter dieser Annahme lassen sich **gleichsetzen**:

- die Gleichung für den Schwierigkeitsindex: $p = N_R/N$
- und die Ausgangsgleichung des Rasch-Modells: $p = e^{\beta-\delta}/(1 + e^{\beta-\delta})$.

Demnach gilt:

$$\frac{N_R}{N} = p = \frac{e^{\beta-\delta}}{1 + e^{\beta-\delta}}$$

In dieser Gleichung sind bekannte und unbekannte Terme:

- **Bekannt** sind N_R und N , somit auch p , der Schwierigkeitsindex.
- **Unbekannt** ist $\beta - d$, die Differenz zwischen Personen- und Itemparameter

Um die unbekannten Terme β und δ zu ermitteln, muß man die Grundgleichung des Rasch-Modells auflösen nach diesen beiden Unbekannten $\beta-d$, man muß also die Inverse von $L(x)$ bilden. Die Auflösung ergibt:

$$L(x)^{-1} = \ln \frac{p}{1-p} = \beta - \delta$$

In Worten: Die Differenz von Personen- und Itemparameter $\beta-d$ läßt sich ermitteln als Logarithmus naturalis aus dem Quotienten des Schwierigkeitsindex p und seines Komplements $(1-p)$.

Die Ermittlung der Schwierigkeitsindizes leitet die Konstruktion einer Rasch-Skala ein.

Die Inverse der Logistischen Funktion wird auch *Logit-Funktion* genannt, ihre Werte heißen *Logits*, sie übernehmen eine Mittlerrolle auf dem Weg einer Rasch-Skalierung.

Verständlich wird diese Skizze des Lösungsweges erst bei Schilderung der einzelnen Konstruktionsschritte.

6.3 Schritte einer Rasch-Skalierung

Der Weg einer Rasch-Skalierung sei exemplarisch an fünf Schritten veranschaulicht:

- (I) : Erstellung einer Matrix
von Schwierigkeitsindizes (Matrix I) (6.3.1),
- (II) : Transformation von Matrix I
in eine Logit-Matrix (Matrix II) (6.3.2),
- (III) : Schätzung der Personen- und Itemparameter
aus Matrix II (Matrix III) (6.3.3),
- (IV) : Reproduktion der Ausgangsmatrix I
aus Matrix III: Modelltest (6.3.4),
- (V) : Standardisierung der ermittelten Personen- und Itemparameter (6.3.5).

6.3.1 Schritt 1:

Erstellung einer Matrix von Schwierigkeitsindizes (Matrix I)

Ermittelt sei eine Matrix von Schwierigkeitsindizes für einen Test. Um den Personenparameter β und den Itemparameter δ getrennt schätzen zu können, nimmt man an, daß in den Schwierigkeitsindizes schon unterschiedliche Personenparameter und unterschiedliche Itemparameter enthalten sind.

Die Annahmen besagen im einzelnen:

- a) Die Matrix der Schwierigkeitsindizes repräsentiert Lösungswahrscheinlichkeiten der Items, in die sowohl Personen- als auch Itemparameter eingehen.
- b) Die Personen lassen sich unterschiedlichen Tüchtigkeitsgruppen zuordnen, die Items unterschiedlichen Schwierigkeitsklassen.

Beispielsweise sei angenommen, daß sich **Personengruppen A, B, C** bilden lassen, die unterschiedliche Personenparameter repräsentieren, und zwar so, daß gilt:

- Gruppe A repräsentiert ‚tüchtige‘ Probanden,
- Gruppe B repräsentiert ‚tüchtigere‘ Probanden,
- Gruppe C repräsentiert ‚noch tüchtigere‘ Probanden.

Ebenso sei angenommen, daß sich **Itemgruppen 1, 2, 3** bilden lassen, die unterschiedliche Itemparameter repräsentieren, und zwar so, daß gilt:

- Item 1 repräsentiert ein ‚leichtes‘ Item,
- Item 2 repräsentiert ein ‚schwereres‘ Item,
- Item 3 repräsentiert ein ‚noch schwereres‘ Item.

Beispiel: 300 Probanden haben einen Test bearbeitet. Sie lassen sich aufteilen in drei Gruppen A, B, C zu je 100 Probanden. Jede Gruppe hat drei Items bearbeitet. Die Lösungshäufigkeiten gibt Kasten 6-4.

Kasten 6-4:

Lösungsmatrix für drei Gruppen A, B, C.

In den Zellen die Häufigkeiten der Richtiglöser (NR) und Falschlöser (NF)

Items	Probanden-Gruppen					
	A		B		C	
	N _R	N _F	N _R	N _F	N _R	N _F
1	71	29	81	19	91	9
2	61	39	71	29	81	19
3	51	49	61	39	71	29
	n _A = 100		n _B = 100		n _C = 100	

Die Gruppierung in Kasten 6-4 dient folgenden drei Schritten:

- Aus der Zahl der Richtiglöser (N_R) und der Zahl der beteiligten Probanden (N) werden die *Schwierigkeitsindizes* p berechnet.
- Aus den Schwierigkeitsindizes wird die *Differenz* $\beta - d$ erschlossen.
- Aus dieser Differenz werden die *Einzelparameter* β und δ isoliert.

Zunächst werden demnach aus den Werten des Kastens 6-4 die Schwierigkeitsindizes ermittelt. Das Ergebnis repräsentiert *Matrix I* in Kasten 6-5. - Der Berechnungsschritt sei an einem Beispiel veranschaulicht. Der Schwierigkeitsindex für ‚Item 1, Gruppe A‘ ($p_{1,A}$) ergibt sich wie folgt:

$$p_{1,A} = \frac{N_{R,A}}{n_A} = \frac{71}{100} = 0.71$$

Den Wert $p_{1,A} = 0.71$ tragen wir ein in die Matrix I, Kasten 6-5: in die Zelle ‚Gruppe A / Item 1‘. - Analog verläuft die Berechnung der anderen Zellen.

Kasten 6-5:

Matrix I - Schwierigkeitsindizes

für drei Probanden-Gruppen A, B, C:		für drei Itemklassen 1, 2, 3:	
A: ‚tüchtige‘ Probanden,		1: ‚leichtes‘ Item,	
B: ‚tüchtigere‘ Probanden,		2: ‚schwereres‘ Item,	
C: ‚noch tüchtigere‘ Probanden;		3: ‚noch schwereres‘ Item.	
Probanden-Gruppen			
Items	A	B	C
1	.71	.81	.91
2	.61	.71	.81
3	.51	.61	.71

Interpretation: Die Schwierigkeitsmatrix in Kasten 6-5 wird interpretiert als *Wahrscheinlichkeitsmatrix*: Jeder Wert p enthält eine ‚Mischung‘ aus *Personenparameter* β und *Itemparameter* δ :

- Von Gruppe A zu C hin nimmt die ‚Tüchtigkeit‘ β zu. Somit werden die Items von A zu C hin eher gelöst, die Lösungswahrscheinlichkeit wächst von A zu C ($p \Rightarrow 1$).

- Von Item 1 zu 3 hin nimmt die ‚Schwierigkeit‘ δ zu. Somit werden die Items von 1 zu 3 hin schwerer lösbar die Lösungswahrscheinlichkeit fällt von Item 1 zu Item 3 ($p \Rightarrow 0$).

Noch einmal ausführlicher:

Jedes einzelne Item (1, 2, 3) repräsentiert dieselbe Schwierigkeit d über alle drei Personen-Gruppen hinweg. Dabei repräsentiert Item 1 eine geringere Schwierigkeit als Item 2 oder 3 ($S_1 < \delta_2 < \delta_3 \Rightarrow p_1 > p_2 > p_3$). In jeder Gruppe A, B, C steigt die Schwierigkeit von Item 1 zu Item 3 (demgemäß fällt p in jeder Gruppe von Item 1 zu 3).

Jede **Gruppe** (A, B, C) repräsentiert dieselbe ‚Tüchtigkeit‘ β über alle Items hinweg. Dabei repräsentiert Gruppe C eine größere Tüchtigkeit als Gruppe B oder A ($\beta_C > \beta_B > \beta_A \Rightarrow p_C > p_B > p_A$). Demnach ist die Wahrscheinlichkeit, ein Item zu lösen, für Gruppe C größer als für Gruppe B oder A (p wächst bei jedem Item von Gruppe A zu C).

Was Matrix I in Kasten 6-5 leisten soll, wird erst aus den weiteren Schritten verständlich. Aber die ‚Leistung‘ sei antizipiert: Matrix I soll so ‚ausgewertet‘ werden, daß (später!) der Personenparameter β für jede Gruppe (A, B, C) und der Itemparameter δ für jedes Item (1, 2, 3) geschätzt werden können und diese ‚Schätzer‘ es dann erlauben, in einem Modelltest die Ausgangswerte zu reproduzieren.

6.3.2 Schritt II:

Transformation von Matrix I in eine Logit-Matrix (Matrix II)

Wir gehen aus von der Gleichung: $p = e^{\text{dif}} / (1 + e^{\text{dif}})$.

- Auf der linken Seite ist der Term p bekannt: Matrix I in Kasten 6-5 gibt die p -Werte an.
- Auf der rechten Seite ist ein Term unbekannt: die Differenz zwischen Personen- und Itemparameter: $\text{dif} = \beta - \delta$.

Um die Differenz ‚dif‘ zu ermitteln, wird die Matrix I (Schwierigkeitsindizes) umgewandelt in eine Logit-Matrix (Matrix II), nach der Inversen von $L(x)$:

$$L(x)^{-1} = \ln \frac{p}{1-p} = \text{dif}$$

Mit Hilfe dieser Gleichung wird aus Matrix I die neue Matrix II ermittelt: die Matrix der Logits - Kasten 6-6 stellt sie dar. Wie die Logits berechnet werden, sollen Beispiele erläutern.

Ermittlung der Logits

Die Zellenwerte (die Logits) von Matrix II (in Kasten 6-6) berechnen sich wie folgt aus $L(x)^{-1} = \text{dif} = \ln [p/(1-p)]$.

HINWEIS: Bei der Berechnung schreiben wir wie bei Statistiken üblich, lateinische Buchstaben: für β das lateinische b , für b das lateinische d .

Item 1, Gruppe A:

Der Schwierigkeitsindex $p_{1,A}$ ist der Matrix I entnommen (Kasten 6-5): $p_{1,A} = 0.71$. Das zugehörige **Logit für die Differenz $\text{dif}_{1,A}$** ergibt sich wie folgt:

$$\text{dif}_{1,A} = \ln [0.71/(1 - 0.71)] = \ln 2.4482 = 0.8953$$

$\text{dif}_{1,A} = \text{Logit}_{1,A} = 0.89$ tragen wir in Matrix II (Kasten 6-6) ein: in die Zelle ‚Item 1/Gruppe A‘.

Item 3, Gruppe B:

Der Wert von p für Item 3 in Gruppe B der Matrix I (Kasten 6-5) beträgt $p_{3,B} = 0.61$. Das zugehörige Logit für die Differenz $(b-d)_{3,B}$ ergibt sich wie folgt:

$$\text{dif}_{3,B} = \ln [0.61/(1 - 0.61)] = \ln 1.5641 = 0.4473$$

$\text{dif}_{3,B} = \text{Logit}_{3,B} = 0.45$ tragen wir in Matrix II (Kasten 6-6) ein: in die Zelle ‚Items 3/Gruppe B‘.

Für die übrigen sieben Zellen der Matrix II (Kasten 6-6) ermitteln wir die Logits in analoger Weise.

Kasten 6-6:

Matrix II: Matrix der in Logits transformierten Schwierigkeitsindizes aus Matrix I

Item	Probanden-Gruppen		
	A	B	C
1	.89	1.45	2.31
2	.45	.89	1.45
3	.04	.45	3.9

Was ist erreicht mit Berechnung der Logits?

Erreicht ist die Kenntnis der Differenz aus Personen- und Itemparameter: **Die Logits (in den Zellen) der Matrix II repräsentieren die Differenzen $\text{dif} = b - d$** . Aus diesen Differenzen sollen die Einzelparameter für Person und Item, b und d , geschätzt werden.

Wie diese Schätzung möglich ist, stellt Schritt III dar.

6.3.3 Schritt III: Schätzung der Personen- und Itemparameter aus Matrix II (Ergebnis \Rightarrow Matrix III)

Gegeben sind in Matrix II (Kasten 6-6) die Differenzen von Personen- und Itemparameter: $\text{dif} = b - d$.

Gesucht sind die getrennten Personen- und Itemparameter, also b und d getrennt.

Die Suche geht so vor sich, daß aus der Logit-Matrix (Kasten 6-6) die Personen- und die Itemparameter geschätzt werden. Bei dieser Schätzung bestehen zuerst gewisse Freiheitsgrade, später reduzieren sich diese Freiheitsgrade sehr schnell.

Der Veranschaulichung wegen sei eine ‚manuelle Berechnung‘ durchgespielt. Ziel der Berechnung ist es, aus den Logits in Matrix II (Kasten 6-6) Randvektoren zu finden, die als Schätzungen der Personen- und Itemparameter (β und δ) gelten können.

Wir geben Matrix II erneut vor: in Kasten 6-7. Die neue Darstellung nennen wir Matrix III-A: In den Zellen stehen die Logits. Am oberen Rand (Gruppe A, B, C) und am linken Rand (Items 1, 2, 3) zeigen Fragezeichen (?) an, worauf sich die Suche richtet: auf den *Randvektor b* (Personenparameter) und den *Randvektor d* (Itemparameter).

Kasten 6-7:
Matrix III-A: Suche nach Personen- und Itemparameter

1. In den Zellen stehen die Logits aus Matrix II (Kasten 6-6).
2. Aus diesen Logits sollen ermittelt werden:
 - Zeile b, die *Personenparameter* und
 - Spalte d, die *Itemparameter*.

Siehe den laufenden Text!

		<i>Probanden-Gruppen</i>			
		b ►	A	B	C
<i>Item</i>	d		b _A ?	b _B ?	b _C ?
1	d ₁ ?		0.89	1.45	2.31
2	d ₂ ?		0.45	0.89	1.45
3	d ₃ ?		0.04	0.45	0.89

Die Logits der Matrix III-A (Kasten 6-7) repräsentieren die Differenzen bd . Demnach müssen die Randvektoren für b und d so ausfallen, daß ihre Differenzen $b-d$ wieder die Logits in den Zellen ergeben.

Der nächste Kasten 6-8 stellt die Ergebnisse dar: Matrix III-B.

Weg von Matrix III-A zu Matrix III-B: Isolierung der Personen- und Itemparameter

Item 1:

Wir gehen aus von Matrix III-A in Kasten 6-7.

In Item 1 stehen unter A: $\text{Logit} = 0.89$, unter B: $\text{Logit} = 1.45$, unter C: $\text{Logit} = 2.31$. Die neuen Werte für b und d (Personen- und Itemparameter) ergeben sich wie folgt:

Gruppe A:

Bei Beginn der Berechnung *nehmen wir an*, daß als Personenparameter in Gruppe A der Wert 1 geeignet sei.

Den Wert $b_A = 1.00$ setzen wir in Matrix III-B (Kasten 6-8) oberhalb von Item 1 unter Gruppe A ein.

Wir haben diesen Wert $b_A = 1.00$ frei gewählt. Nun gilt die Gleichung:

$$b_A - d_1 = \text{dif}_{1,A}.$$

Von den drei Größen dieser Gleichung sind zwei bekannt:

1. $\text{dif}_{1,A} = \text{Logit}_{1,A} = 0.89$,
2. $b_A = 1.0$ *Soeben festgelegt!*

Einsetzen in die Gleichung ergibt: $1.00 - d_1 = 0.89 \Rightarrow d_1 = 0.11$

Den Wert $d_1 = 0.11$ tragen wir in Matrix III-B (Kasten 6-8) neben Item 1 ein.

Aus der Festlegung $b_A = 1$ und $d_1 = 0.11$ ergeben sich die weiteren Werte für Gruppe B und C, also zwei Personenparameter.

Gruppe B:

Zu ermitteln ist b_B . Gegeben ist $d_1 = 0.11$. Das Ergebnis der Differenz von b_B und d_1 ist festgelegt in der Zelle von 'Item 1, Gruppe B' mit $\text{Logit}_{1,B} = \text{dif}_{1,B} = 1.45$.

$$\text{Es folgt: } b_B - 0.11 = 1.45 \Rightarrow b_B = 1.45 + 0.11 \Rightarrow b_B = 1.56$$

Den Wert $b_B = 1.56$ tragen wir in Matrix III-B (Kasten 6-8) unter Gruppe B ein.

Gruppe C:

Zu ermitteln ist b_C . Gegeben sind $d_1 = 0.11$ und $\text{dif}_{1,C} = \text{Logit}_{1,C} = 2.31$. Es folgt: $b_C = 2.42$.

Den Wert $b_C = 2.42$ tragen wir in Matrix III-B (Kasten 6-8) unter Gruppe C ein.

Erstes Zwischenergebnis: Ermittelt haben wir

- drei Personenparameter: $b_A = 1.00$, $b_B = 1.56$, $b_C = 2.42$ und
- einen Itemparameter: $d_1 = 0.11$.

Item 2:

Wir gehen erneut aus von Matrix III-A in Kasten 6-7.

Bei Item 2 stoßen wir an die Grenzen unserer manuellen Berechnung: Wir kommen nicht mehr zu gleichen Itemparametern, wenn wir aus den Parametern b_A , b_B , b_C und d_1 , die wir schon ermittelt haben, die Parameter für Item 2 und Item 3 schätzen wollen.

Um aber den Gedankengang weiter zu veranschaulichen (auch, um die Grenzen des vereinfachten Vorgehens sichtbar zu machen), führen wir die Berechnungsschritte fort.

Bekannt bei Item 2 sind zwei Sachverhalte:

1. Bekannt sind die Differenzen $b-d$: die Logits in den Zellen,
 Item 2, Gruppe A: $\text{dif}_{2,A} = 0.45$,
 Item 2, Gruppe B: $\text{dif}_{2,B} = 0.89$,
 Item 2, Gruppe C: $\text{dif}_{2,C} = 1.45$.
2. Bekannt sind auch die zugehörigen Personenparameter nämlich aus der Berechnung von b und d für Item 1:
 $b_A = 1.00$,
 $b_B = 1.56$,
 $b_C = 2.42$.

Gesucht wird der Itemparameter d_2 .

Gruppe A:

Gegeben ist $\text{dif}_{2,A} = 0.45$ und $b_A = 1.00$. Gesucht ist d_2 . Für d_2 gilt: $\text{dif}_{2,A} = b_A - d_2 \Rightarrow 0.45 = 1.00 - d_2 \Rightarrow d_2 = 0.55$.

Gruppe B:

Gegeben ist $\text{dif}_{2,B} = 0.89$ und $b_B = 1.56$. Gesucht ist d_2 . Somit gilt: $0.89 = 1.56 - d_2 \Rightarrow d_2 = 0.67$.

Der Wert für d_2 , ermittelt in Gruppe B, stimmt nicht überein mit dem Wert, der für d_2 in Gruppe A ermittelt wurde.

Gruppe C:

Gegeben ist $\text{dif}_{2,C} = 1.45$ und $b_C = 2.42$. Gesucht ist erneut $d_2 \Rightarrow d_2 = 0.97$.

Bei Gruppe C ergibt sich ein anderer Wert für d_2 als in Gruppe A und in Gruppe B.

Als Kompromiß sei das arithmetische Mittel der drei Schätzwerte berechnet: $(0.55 + 0.67 + 0.97)/3 = d_2 = 0.73$.

Diesen Mittelwert $d_2 = 0.73$ setzen wir in die Matrix III-B (Kasten 6-8) neben Item 2 ein.

Item 3:

Wieder gehen wir aus von Matrix III-A in Kasten 6-7.

Von Item 3 sind ebenfalls zwei Sachverhalte *bekannt*:

1. die Differenzen bd : die Logits von
Item 3, Gruppe A: $\text{dif}_{3,A} = 0.04$,
Item 3, Gruppe B: $\text{dif}_{3,B} = 0.45$,
Item 3, Gruppe C: $\text{dif}_{3,C} = 0.89$;
2. die zugehörigen *Personenparameter*
 $b_A = 1.00$,
 $b_B = 1.56$,
 $b_C = 2.42$.

Für **Gruppe A** folgt: $d_3 = 0.96$.
Für **Gruppe B** folgt: $d_3 = 1.11$.
Für **Gruppe C** folgt: $d_3 = 1.53$.

Das arithmetische Mittel der drei Schätzwerte, $d_3 = 1.20$, setzen wir in Matrix III-B (Kasten 6-8) neben Item 3 ein.

Zweites Zwischenergebnis: Wir haben zwei weitere Itemparameter ermittelt: $d_2 = 0.73$ und $d_3 = 1.20$. Zum Zeichen, daß diese beiden Werte Kompromisse repräsentieren, setzen wir hinter ihre Angabe ein Fragezeichen in Kasten 6-8.

Kasten 6-8 repräsentiert nun das Ergebnis einer Isolierung von Item- und Personenparameter.

Kasten 6-8:
Matrix III-B:

Trennung von Personenparameter (Zeile: b) und Itemparameter (Spalte: d)

Neu gegenüber Matrix III-A (in Kasten 6-7) sind die geschützten Rundvektoren <i>b</i> und <i>d</i> .					
Zur Berechnung siehe den laufenden Text!					
		Probanden-Gruppen			
			A	B	C
<i>Item</i>	d	b ►	1.00	1.56	2.42
1	0.11		0.89	1.45	2.31
2	0.73 ?		0.45	0.89	1.45
3	1.20 ?		0.04	0.45	0.89

Resümee zu Schritt III: Wenn wir die Personen- und Itemparameter ermittelt haben, dann ist es in der Tat gelungen, aus den bekannten Werten p (Schwierigkeitsindizes) die unbekannten Größen b und d *getrennt* zu schätzen.

6.3.4 Schritt IV - ein Modelltest:

Reproduktion der Ausgangsmatrix Matrix I

Uns liegen vor die geschätzten Werte für Item- und Personenparameter in Matrix III-B (Kasten 6-8). Doch erhebt sich eine neue Frage: Sind die ermittelten Personen- und Itemparameter auch zutreffend?

Zutreffend sind sie, wenn sie es erlauben, die Ausgangsmatrix (Matrix I: Schwierigkeitsindizes) zu reproduzieren. Somit stehen wir vor der Aufgabe, einen *Modelltest* durchzuführen:

- Aus den Personenparametern (Zeile **b**) und den Itemparametern (Spalte **d**) der Matrix III-B in Kasten 6-8
- sollen die Werte p der entsprechenden Zellen der Matrix I in Kasten 6-5 reproduziert werden.

Diesen Modelltest beschreiben wir in Schritt IV. Der Modelltest besteht in einem Vergleich von Matrix III-B (in Kasten 6-8) und Matrix I (in Kasten 6-5).

Der Vergleich soll prüfen, ob

- die **geschätzten** Parameter der Matrix III-B,
 - die **originalen** Schwierigkeitsindizes der Matrix I
- hinreichend genau reproduzieren.

Bei hinreichend genauer Reproduktion gelten die Personen- und Itemparameter der Matrix III-B als verträglich mit dem Raschmodell und in diesem Sinne als ‚bestätigt‘.

Vorbereitung des Modelltests

Bei dem Vergleich ist auszugehen von der Logistischen Funktion in ihrer Anwendung auf die Fragestellung:

$$L(x) = \frac{e^x}{1 + e^x} = \frac{e^{\beta - \delta}}{1 + e^{\beta - \delta}} = p$$

In die Logistische Gleichung setzen wir für β und δ die geschätzten Werte der Matrix III-B ein.

Um diesen Schritt zu veranschaulichen, geben wir zweierlei vor:

1. Die Werte für b und d aus Matrix III-B (Kasten 6-8) tragen wir in eine neue Matrix ein, die wir Matrix IV-A nennen (Kasten 6-9).
2. Die gesuchten Schwierigkeitsindizes zeigen wir an, indem wir in die einzelnen Zellen der Matrix IV-A ein p schreiben und mit einem Fragezeichen versehen.

Kasten 6-9:**Matrix IV-A: Modelltest: Reproduktion der Matrix I aus Matrix III-B**

1. Matrix IV-A enthält die Randvektoren b und d aus Matrix III-B.
2. Gesucht sind die Zellenwerte p: gekennzeichnet als „p?“.

		Probanden-Gruppe		
		A	B	C
Item	d	b ►		
1	0.11	1.00	1.56	2.42
2	0.73	p?	p?	p?
3	1.20		p?	p?

Reproduktion der Schwierigkeitsindizes aus den geschätzten Personen- und Itemparametern

Wie die Schwierigkeitsindizes ermittelt werden, sei an einigen Beispielen demonstriert.

Item 1:

Wir gehen aus von b und d in Matrix IV-A (Kasten 6-9).

Für **Gruppe A** sind gegeben: $b_A = 1.00$ und $d_1 = 0.11$. Einsetzen in die Logistische Funktion ergibt:

$$p_{1,A} = \frac{e^{\beta(A)-\delta(1)}}{1 + e^{\beta(A)-\delta(1)}} = \frac{e^{1.00-0.11}}{1 + e^{1.00-0.11}} = \frac{2.452}{1 + 2.4531} = 0.7088$$

Diesen Wert $p_{1,A} = 0.71$ setzen wir in eine neue Matrix ein, Matrix IV-B (Kasten 6-10): in die Zelle „Item 1, Gruppe A“.

Für **Gruppe B** berechnet sich der Zellenwert wie folgt:

$$p_{1,B} = \frac{e^{1.56-0.11}}{1 + e^{1.56-0.11}} = \frac{4.2631}{1 + 4.2631} = 0.8099$$

Den Wert $p_{1,A} = 0.81$ tragen wir in Matrix IV-B (Kasten 6-10) ein: in die Zelle „Item 1, Gruppe B“.

Für **Gruppe C** ergibt sich: $p_{1,C} = 0.9097$. Den Wert $p_{1,C} = 0.91$ tragen wir in Matrix IV-B (Kasten 6-10) ein: in die Zelle „Item 1, Gruppe C“.

Ergebnis für Item 1: Bei allen drei Gruppen A, B, C gelingt die Reproduktion der Schwierigkeitsindizes exakt - natürlich, die Zahlen sind ja systematische Ableitungen aus Matrix I, II und III.

Item 2:

Wir gehen erneut aus von b und d in Matrix IV-A (Kasten 6-9).

Die Reproduktion gelingt keineswegs mehr so exakt wie für Item 1 - weil ja der Itemparameter d_2 „von Hand“ nicht eindeutig zu bestimmen war. Wir hatten, zur Demonstration, einen mittleren Wert von $d_2 = 0.73$ eingetragen.

Für **Gruppe A, Gruppe B und Gruppe C** werden folgende p-Werte reproduziert:

$$\begin{aligned} p_{2,A} &= 0.5671, \\ p_{2,B} &= \mathbf{0.6963}, \\ p_{2,C} &= \mathbf{0.8442}. \end{aligned}$$

Ergebnis für Item 2: Bei keiner Gruppe A, B, C fällt die Reproduktion der Schwierigkeitsindizes exakt aus, aber für jede noch annähernd genau.

Item 3:

Wieder gehen wir aus von b und d in Matrix IV-A (Kasten 6-9).

Die Reproduktion bleibt ebenso ungenau wie für Item 2 - weil auch Itemparameter d_3 manuell nicht eindeutig zu bestimmen war. Wir hatten als mittleren Wert $d_3 = 1.20$ eingetragen.

Für Gruppe A, B, C werden folgende p-Werte reproduziert:

$$\begin{aligned} p_{3,A} &= \mathbf{0.4502}, \\ p_{3,B} &= \mathbf{0.5866}, \\ p_{3,C} &= 0.7721. \end{aligned}$$

Ergebnis für Item 3: Die Rückrechnung erbringt ungenaue Werte. Die reproduzierten p-Werte decken sich nicht mit den ursprünglichen Schwierigkeitsindizes der Matrix I (Kasten 6-5).

Matrix IV-B in Kasten 6-10 repräsentiert das Ergebnis: Aus Matrix IV-A wurden Wahrscheinlichkeitsindizes p reproduziert, die den Schwierigkeitsindizes in Matrix I (Kasten 6-5) entsprechen sollen.

Kasten 6-10:

Matrix IV-B: Modelltest: Reproduktion der Matrix I aus Matrix III-B

Ermittelt werden die Zellenwerte p aus den Randvektoren b und d.					
<i>Zur Berechnung siehe den laufenden Text!</i>					
		Probanden-Gruppe			
		<div style="display: flex; justify-content: space-around;"> A B C </div>			
Item	d	b ►	1.00	1.56	2.42
1	0.11		0.71	0.81	0.91
2	0.73		0.57 ?	0.70 ?	0.84 ?
3	1.20		0.45 ?	0.58 ?	0.77 ?

Durchführung des Modelltests

Ob die reproduzierten Schwierigkeitsindizes in Matrix IV-A (Kasten 6-10) von den originalen Schwierigkeitsindizes in Matrix I (Kasten 6-5) erheblich abweichen, prüfen wir mit dem χ^2 -Test; die Formel lautet:

$$\chi^2 = \sum \frac{(f_b - f_e)^2}{f_e}$$

E_s bedeuten:

χ^2 : Prüfgröße Chi-Quadrat,

f_b : Häufigkeit der beobachteten Werte,

f_e : Häufigkeit der erwarteten Werte.

Die Freiheitsgrade (FG) berechnen sich nach der Formel:

$$FG = (n_{\text{Reihe}} - 1) \cdot (n_{\text{Spalte}} - 1)$$

Es bedeuten:

n_{Reihe} : Zahl der Reihen,

n_{Spalte} : Zahl der Spalten.

Absolute Häufigkeiten: Der χ^2 -Test bezieht sich auf absolute Häufigkeiten. In Matrix IV-B (Kasten 6-10) stehen relative Häufigkeiten (p). Wir wandeln sie um in absolute Häufigkeiten nach der Formel: $p = N_R/N \Rightarrow N_R = p \cdot N$.

N_R bezeichnet die Zahl der Richtiglöser, N die Zahl der beteiligten Probanden.

Wir geben ein Beispiel: In Matrix IV-B (Kasten 6-10) beträgt p für Item 3 in Gruppe B: $p_{3,B} = 0.58$. N für Gruppe B beträgt $N = 100$ (bekannt aus Kasten 6-4: $n_B = 100$). Daraus folgt:

$$N_{R,3,B} = 0.58 \cdot 100 = 58$$

Ähnlich berechnen sich die Häufigkeiten der anderen acht Zellen.

Wir tragen die errechneten Werte in Kasten 6-11 ein und nennen die neue Matrix IV-C.

Wir tun dabei einen weiteren Schritt: In Kasten 6-11 fügen wir die absoluten Werte aus Kasten 6-4 für N_R hinzu. Damit stellen wir in Kasten 6-11 für jede Gruppe A, B, C zwei Häufigkeitsspalten nebeneinander:

- die originalen Werte für N_R aus Kasten 6-4 und
- die reproduzierten Werte für N_R , ermittelt aus Kasten 6-10.

Für die statistische Prüfung betrachten wir

- die Originalwerte N_R als die erwarteten Werte: f_e ,
- die reproduzierten Werte N_R als die beobachteten Werte: f_b .

Kasten 6-11:
Matrix IV-C: Modelltest mittels Chi-Quadrat-Test

Verglichen werden je Gruppe A, B, C zwei Häufigkeitsreihen:

- f_a : Originalwerte für Richtiglöser N_R aus Kasten 6-4,
- f_b : reproduzierte Werte für Richtiglöser N_R aus Kasten 6-10.

Zum Verständnis siehe den laufenden Text!

Items	Probanden-Gruppen					
	A		B		C	
	f_a	f_b	f_a	f_b	f_c	f_b
1	71	71	81	81	91	91
2	61	57	71	70	81	84
3	51	45	61	58	71	77

$$X^2 = 1.82 \text{ bei } FG = 4$$

Entscheidung: Der empirische X^2 -Wert = 1.82 bei $FG = 4$ liegt mit $p < 1$ Prozent unterhalb des kritischen Wertes $X^2 = 13.27$. Demnach unterscheiden sich die reproduzierten Werte in Kasten 6-11 nicht signifikant von den originalen Werten in Kasten 6-4.

Demnach können die errechneten Personen- und Itemparameter - im Sinne einer Demonstration - als ‚angemessene‘ Schätzungen von ‚Tüchtigkeit‘ und ‚Itemschwierigkeit‘ gelten.

4.3.5 Standardisierung der ermittelten Personen- und Itemparameter

Die Matrix III-B in Kasten 6-8 liefert die gesuchten Informationen über Item- und Personenparameter. Doch sind die Personen- und Itemparameter b und d vergleichsweise ‚unpraktisch‘ ausgefallen - wenig geeignet für eine Übersicht:

$$\begin{array}{ll} b_A = 1.00, & d_1 = 0.11, \\ b_B = 1.56, & d_2 = 0.73, \\ b_C = 2.42, & d_3 = 1.20. \end{array}$$

Um den Nachteil der Unübersichtlichkeit auszugleichen, ist ein weiterer Schritt sinnvoll: eine Standardisierung der ermittelten Personen- und Itemparameter.

Beispielsweise kann zu allen Itemparametern eine Konstante hinzutreten, so daß der Wert für die mittlere Schwierigkeit bei $d_m = 0$ liegt (oder bei einer anderen gewünschten Maßzahl).

„In der Regel wird die Skala für die Itemschwierigkeit so normiert, daß Itemparametersummen von 0 resultieren; in der Mehrzahl der Fälle liegen dann die Parameter d_i im Intervall zwischen -3 und +3“ (Michel & Conrad, 1982, 31).

Wir wenden diesen Vorschlag auf unser Übungsbeispiel an und transformieren zuerst die Item-, dann die Personenparameter.

Transformation der Itemparameter: Zu ermitteln ist das mittlere d_m :

$$d_m = (0.11 + 0.73 + 1.20)/3 = 0.68$$

Um den Betrag $d_m = 0.68$ vermindern wir alle drei Itemparameter. Es ergibt sich:

$$d_1 = 0.11 - 0.68 = -0.57,$$

$$d_2 = 0.73 - 0.68 = 0.05,$$

$$d_3 = 1.20 - 0.68 = 0.52.$$

Die Summe der transformierten Werte ergibt Null. Die neuen Werte setzen wir in Kasten 6-12 ein: Spalte d.

Transformation der Personenparameter: Die standardisierten Personenparameter berechnen wir nach der Inversen der Logistischen Funktion: $L(x)^{-1} = \beta - \delta = \ln [p/(1-p)]$. Aufgelöst nach β bzw. nach b , ergibt sich:

$$b = d + \ln [p/(1-p)].$$

In die Berechnung des Personenparameter b_A seien alle drei Itemparameter d_1 , d_2 und d_3 einbezogen:

$$b_A = d_1 + \ln [p_{1,A}/(1 - p_{1,A})] = -0.57 + \ln [0.71/(1 - 0.71)] = 0.33$$

$$b_A = d_2 + \ln [p_{2,A}/(1 - p_{2,A})] = 0.05 + \ln [0.57/(1 - 0.57)] = 0.33$$

$$b_A = d_3 + \ln [p_{3,A}/(1 - p_{3,A})] = 0.45 + \ln [0.45/(1 - 0.45)] = 0.32$$

Nach der Berechnung gilt: $b_A = 0.33$,

entsprechend folgt für b_B und b_C : $b_B = 0.87$,

$$b_C = 1.72.$$

Die neuen Parameter setzen wir in Kasten 6-12 ein: Zeile b.

Kasten 6-12:

Standardisierte Personen- und Itemparameter von Matrix IV-B (Kasten 6-10):

Die Summe von d ergibt Null.

Zum Verständnis siehe den laufenden Text!					
			Probanden-Gruppe		
			A	B	C
Item	d	b ►	0.33	0.87	1.72
1	-0.57		0.71	0.81	0.91
2	0.05		0.57	0.70	0.84
3	0.52		0.45	0.58	0.77

Zum Ergebnis der Rasch-Skalierung

Unsere Berechnung hat *nur einer Demonstration* gedient. Hätten wir stattdessen einen ‚echten probabilistischen Test‘ konstruiert, dann stünde nun ein In-

strument zur Verfügung, mit dem sich Probanden, die der modellkonformen Population entsprechen, genau einstufen ließen nach ihrer Tüchtigkeit.

Darüber hinaus könnte es sinnvoll sein, den probabilistischen Test Prozeduren zu unterwerfen, die zum Repertoire der klassischen Testtheorie gehören. Gedacht ist vorrangig an zwei Maßnahmen:

- Für den ‚neuen‘ Test ließen sich die klassischen Gütekriterien (Objektivität, Reliabilität und Validität) bestimmen.
- Vor allem ließe sich eine klassische Eichung vornehmen, die dann Normen für die Gesamtgruppe und für unterschiedliche Teilgruppen erbringen könnte (vgl. Kubinger & Wurst, 1994).

6.4 Ergänzende Hinweise zur Endmatrix

Zur Endmatrix einer Rasch-Skala lassen sich Hinweise geben, welche die Ergebnisse einem umfassenderen Kontext zuordnen. Verwiesen sei auf

- die Ermittlung von Vertrauensbereichen (6.4.1),
- die Berechnung von Standardmeßfehlern (6.4.2),
- die iterative Berechnung von Modellparametern (6.4.3),
- effektivere Algorithmen zur Parameterschätzung (6.4.4).

6.4.1 Ermittlung von Vertrauensbereichen

Für eine Raschskala lassen sich Vertrauensbereiche berechnen (Fricke, 1974, 94). Diese Möglichkeit sei demonstriert für die standardisierte Endmatrix in Kasten 6-12. Die Summe der Itemparameter ergibt den Wert Null. Die Inverse der Logistischen Funktion, aufgelöst nach β , lautet für den Fall ($\Sigma\delta$) = 0:

$$\beta = \ln \frac{p}{1-p} + \delta = \ln \frac{p}{1-p} + 0$$

Zwei Beispiele:

1. Wie groß ist - bei der standardisierten Matrix - der Vertrauensbereich bei $p = 0.90$? Einsetzen der Werte in die nach β bzw. b aufgelöste Irrverse ergibt:

$$b = \ln [0.90/(1 - 0.90)]$$

$$b = 2.1972$$

Interpretation: „Personen werden . . . 90 % der Aufgaben richtig lösen, wenn sie einen Fähigkeitsparameter von $b = 2.1972$ aufweisen“ (Fricke, 1974, 94).

2. Wie groß ist - bei der standardisierten Matrix - der Vertrauensbereich bei $p = 0.95$? Einsetzen ergibt:

$$b = \ln [0.95/(1 - 0.95)]$$

$$b = 2.94$$

Interpretation: Personen, die einen Personenparameter von $b = 2.94$ haben, werden wahrscheinlich 95 Prozent der Aufgaben richtig lösen.

6.4.2 Berechnung von Standardmeßfehlern

Wright and Stone (1979, 22) geben Algorithmen an, die es erlauben, Standardmeßfehler (SE) zu schätzen:

- (1) für Personen- und
- (2) für Itemparameter

(1) SE für Personenparameter

Der Standardmeßfehler für den Personenparameter berechnet sich nach der Formel:

$$SE_{\beta,i} = g_{\beta} \cdot \sqrt{\frac{n_{it}}{nr_{it,i} \cdot (n_{it} - nr_{it,i})}}$$

Es bedeuten:

- $SE_{\beta,i}$: Standardmeßfehler eines Probanden mit dem Personenparameter $b(i)$,
 g_{β} : Gewichtungsfaktor für b (≈ 2.10),
 n_{it} : Zahl der Testitems,
 $nr_{it,i}$: Zahl der Items, die ein Proband mit $\beta(i)$ richtig gelöst hat.

Beispiel: Ein Rasch-skaliertes Test habe $n_{it} = 14$ Items. Proband 12 beantwortete $nr_{it,12} = 1$ Item richtig (Wright & Stone, 1979, 43). Dann beträgt der Standardmeßfehler dieses Probanden: $SE_{\beta,rz} = 2.18$.

$$SE_{\beta,12} = 2.10 \cdot \sqrt{\frac{14}{1 \cdot (14 - 1)}} = 2.10 \cdot 1.04 = 2.18$$

(2) SE für Itemparameter

Der Standardmeßfehler für den Itemparameter berechnet sich nach der Formel:

$$SE_{\delta,j} = g_{\delta} \cdot \sqrt{\frac{N_j}{NR_j \cdot (N_j - NR_j)}}$$

Es bedeuten:

- $SE_{\delta,j}$: Standardmeßfehler des Items j ,
 g_{δ} : Gewichtungsfaktor für δ (≈ 1.31),
 N_j : Zahl der Probanden, die Item j beantwortet haben,
 NR_j : Zahl der Probanden, die Item j richtig gelöst haben.

Beispiel: Das Item 1 eines Rasch-skalierten Tests werde beantwortet von $N_I = 34$ Probanden. Davon mögen $N_{RI} = 32$ eine richtige Antwort geben (Wright & Stone, 1979, 41). Dann beträgt der Standardmeßfehler: $SE_{s,I} = 0.95$.

$$SE_{\delta,1} = 1.31 \cdot \sqrt{\frac{34}{32 \cdot (43 - 32)}} = 1.31 \cdot 0.73 = 0.95$$

6.4.3 Iterative Berechnung von Modellparametern

Wir haben in unserem Beispiel die beiden Parameterklassen β und δ nur *einmal* berechnet und dann einem Modelltest unterworfen. Doch braucht der Testautor sich nicht mit einer *einmaligen* Schätzung zu begnügen. Beide Parameter kann er zu wiederholten Malen berechnen. Die zuerst ermittelten Personen- und Itemparametern dienen dann dazu, die Parameter erneut zu schätzen und sie erneut einem Modelltest zu unterwerfen.

Es sei nur die Grundidee skizziert.

Beispiel in drei Schritten:

1. Wir gehen aus von den drei Personenparametern β_A, β_B und β_C die wir in unserer Demonstration geschätzt haben. (Wir vergessen gleichsam die Itemparameter)
2. Wir berechnen neue Itemparameter δ_1, δ_2 und δ_3 ; mit Hilfe der Inversen der Logistischen Funktion, nach der gilt: $\delta = \beta - \ln [p/(1-p)]$.
3. Wir setzen die neu geschätzten Parameter β und δ in die Logistische Funktion ein (wie schon einmal geschehen in Schritt IV, S. 167) und prüfen, ob die auf diese Weise neu ermittelten Wahrscheinlichkeitsindizes p die Matrix I (in Kasten 6-5) genauer reproduzieren.

Solche Berechnungen lassen sich - im Grunde - beliebig oft wiederholen. Doch keineswegs ist gewährleistet, daß die Iterationen auch verbesserte Parameter liefern.

6.4.4 Effektivere Algorithmen zur Parameterschätzung

Unsere ‚manuelle‘ Schätzung der Parameter β und δ hatte lediglich den Zweck, den Grundgedanken einer Raschskalierung zu veranschaulichen. Die Methode ist ineffizient, wenn es darum geht, größere Datenmengen zu verarbeiten. Dafür stehen andere Schätz- und Kontrollverfahren zur Verfügung.

Ein effektives Näherungsverfahren zur ‚manuellen‘ Parameterschätzung beschreiben Wright und Stone (1979, 28-45). Die Berechnung ist aber so auf-

wendig, daß sie den Raum überschreitet, den dieses ‚diagnostische Lehrbuch‘ gewährt.

„Die Parameterschätzung muß in allen Fällen ... unter Verwendung elektronischer Datenverarbeitungsanlagen erfolgen, falls eine für praktische Zwecke ausreichende Itemzahl für den Test konstruiert wurde“ (Wottawa, 1980, 59).

Schätzalgorithmen beschreiben beispielsweise: Andersen (1982), Fischer (1974), Rasch (1960), Scheiblechner (1971), Wright (1977), Wright und Masters (1982), Wright und Stone (1979).

6.5 Charakteristika einer Rasch-Skala

Entspricht ein Itemsatz den Annahmen des Rasch-Modells, dann ergeben sich spezifische Charakteristika. Davon seien besprochen:

- die Homogenität der Items (6.5.1),
- die lokale stochastische Unabhängigkeit der Items (6.5.2),
- die Stichprobenunabhängigkeit von Skala und Items (spezifische Objektivität, Teilgruppenkonstanz) (6.5.3),
- die Separierbarkeit von Item- und Personenparameter (6.5.4).

4.5.1 Homogenität

Gemäß den Modellannahmen werden nur solche Items zugelassen, deren Item-Charakteristik-Kurven (ICC) gleichartig verlaufen und sich nicht schneiden.

Solche Items heißen homogen. Dies soll Kasten 6-13 veranschaulichen.

„Alle Items sind homogen in dem Sinne, daß sich ihre Charakteristiken gleichen bis auf Translationen in Abhängigkeit von den Itemschwierigkeiten“ (Michel & Conrad, 1982, 28). Noch einmal: Alle Items zeigen den gleichen Verlauf der Lösungswahrscheinlichkeiten p ; ihre Kurven verlaufen an unterschiedlichen Stellen auf dem Item-Personenparameter-Kontinuum - gemäß der unterschiedlichen ‚Schwierigkeit‘, die ein Item repräsentiert.

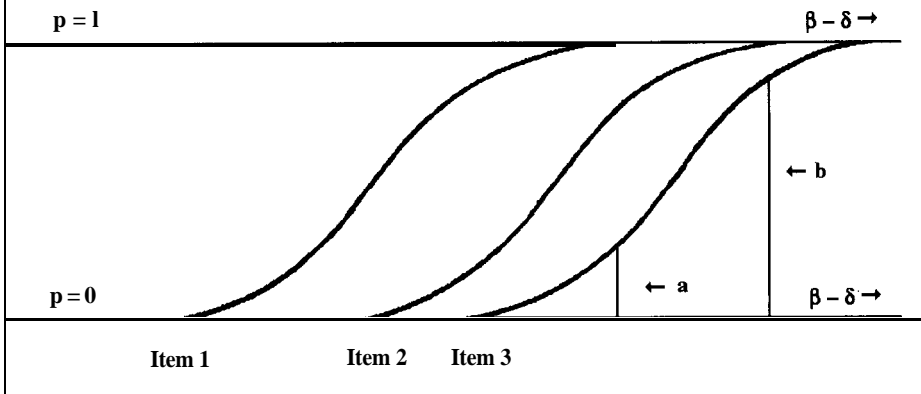
Kasten 6-13:
Homogene Itemcharakteristiken

Homogen sind Items, deren Rem-Charakteristik-Kurven gleichartig verlaufen und sich nicht überschneiden.

- (1) Die drei Kurven repräsentieren die Lösungswahrscheinlichkeiten p der Items 1, 2, 3. Bei jedem Item wächst die Wahrscheinlichkeit ‚nach rechts hin‘, die Ordinate zeigt die **Höhe** der Lösungswahrscheinlichkeit an.

So signalisiert die Ordinate b bei Item 3 eine größere Tüchtigkeit (β) und damit eine höhere Lösungswahrscheinlichkeit als die Ordinate a bei demselben Item 3.

- (2) Items 1, 2, 3 haben unterschiedliche Schwierigkeiten (δ): Item 3 ist schwerer als Item 2 und 1. Zur Lösung setzt Item 3 demgemäß eine größere ‚Tüchtigkeit‘ (β) voraus als Item 2 und 1. In diesem Sinne ‚wächst‘ die Tüchtigkeit ‚von links nach rechts hin‘ an.



Erläuterung zu Kasten 6-13: Der Kurvenverlauf zeigt an, daß mit Wachsendem Personenparameter die Wahrscheinlichkeit zunimmt, das Item zu lösen. Entspricht ein Satz von Items solchen Anforderungen, dann gilt:

- Greifen wir zwei **Personen** heraus, deren Tüchtigkeiten sich unterscheiden, so gilt: Für die ‚tüchtigere‘ Person ist die Wahrscheinlichkeit immer höher, ein Item zu lösen, als für die ‚weniger tüchtige‘ Person.

Veranschaulicht am Kurvenverlauf. Eine ‚Tüchtigkeit‘, die rechts platziert ist, hat die höhere Wahrscheinlichkeit, irgendein Item zu lösen, als eine ‚Tüchtigkeit‘, die links platziert ist.

- Greifen wir zwei **Items** heraus, deren Schwierigkeiten sich unterscheiden, so gilt: Für ein und dieselbe Person ist die Wahrscheinlichkeit immer höher, das Item zu lösen, das ‚leichter‘ ist im Vergleich zu ihrer Tüchtigkeit, als das Item, das ‚schwerer‘ ist.

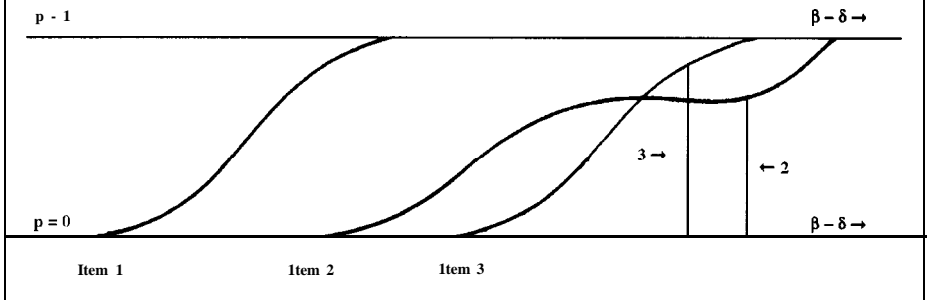
Veranschaulicht am Kurvenverlauf. Ein und dieselbe ‚Tüchtigkeit‘ besitzt eine höhere Wahrscheinlichkeit, ein Item zu lösen, das links von ihr platziert ist, als ein Item, das rechts von ihr platziert ist.

Inhomogen ist ein Item, wenn seine Item-Charakteristik-Kurve die Kurve eines anderen Items schneidet. Kasten 6-14 veranschaulicht einen solchen Verlauf.

Kasten 6-14:
Inhomogene Itemcharakteristiken

- Die drei Item-Charakteristik-Kurven veranschaulichen den Verlauf der Lösungswahrscheinlichkeiten von Item 1, 2 und 3.
- Die Ordinaten zeigen die **Höhe** der Lösungswahrscheinlichkeiten an.
- Nun zeigt sich: Die Kurve von Item 2 schneidet die von Item 3. Rechts der Schnittstelle gilt: Ordinate 3 ist höher als Ordinate 2; das bedeutet: Die Wahrscheinlichkeit, Item 3 zu lösen, ist größer als die, Item 2 zu lösen - entgegen den Forderungen des Raschmodells. Item 2 ist inhomogen.

Zum Ganzen siehe den laufenden Text!



Erläuterung zu Kasten 6-14: ICC 1 ist links von ICC 2 platziert, d.h. Item 1 ist leichter als Item 2.

ICC 2 ist links von ICC 3 platziert, d.h. Item 2 ist leichter als Item 3, es sollte durchgängig leichter sein. **Item 2 ist jedoch nicht durchgängig leichter** - dies zeigt sich an den zugehörigen Ordinaten,

Die Ordinaten zeigen jeweils die Höhe der Lösungswahrscheinlichkeiten an. ICC 2 schneidet ICC 3. Rechts dieser Schnittstelle ist die Ordinate für Item 3 höher als die Ordinate für Item 2; dies bedeutet, daß die Wahrscheinlichkeit, Item 3 zu lösen, größer ist als die, Item 2 zu lösen. Diese ‚Unordnung‘ verstößt gegen die Anforderungen des Rasch-Modells, nach denen ein ‚vorhergehendes‘ Item durchgängig leichter zu lösen sein muß als ein ‚nachfolgendes‘. Item 2 ist inhomogen. Inhomogene Items werden eliminiert.

6.5.2 Lokale stochastische Unabhängigkeit der Items

Es ist „wünschenswert, wenn die Zahl der gelösten Aufgaben die gesamte in den Antworten enthaltene Information bezüglich des Personenparameters enthält, es also keine Rolle spielt, welche Items gelöst wurden, sondern nur, wieviele“ (Wottawa, 1980, 55).

Angezielt wird demnach ein Test, bei dem die Items eine ‚erschöpfende Statistik‘ liefern, eine ‚erschöpfende Information‘ über die ‚Tüchtigkeit‘ eines Probanden. Jede Teilmenge von Items soll die Ausprägung der latenten Dimension erschöpfend charakterisieren.

Um dies zu gewährleisten, müssen die Items ‚lokal stochastisch unabhängig‘ voneinander sein (Michel & Conrad, 1982, 28-29; Wottawa, 1980, 54-55). Dies bedeutet:

- Die Wahrscheinlichkeit, das eine Item zu lösen, darf nicht abhängen von der Wahrscheinlichkeit, das andere Item zu lösen. Anders gesagt: Ein Item gelöst zu haben soll nicht die Wahrscheinlichkeit erhöhen, auch ein anderes Item zu lösen.
- Vielmehr soll gelten: Die Lösung von Item i und die von Item j hängen allein von der Tüchtigkeit des Probanden und der Schwierigkeit des Items ab.

Anders gesagt: Lokale stochastische Unabhängigkeit besagt, daß die Eintretenswahrscheinlichkeit mehrerer Ereignisse nur abhängt von der Eintretenswahrscheinlichkeit jedes einzelnen Ereignisses und nicht von ihrer Kombination.

Beispiel: Betrachtet seien mehrere Würfe mit einem Würfel. Daß jemand bei **einem** Wurf eine Sechs gewürfelt hat, ‚bewirkt‘ nicht, daß er auch bei einem anderen Wurf eine Sechs würfelt. Die gewürfelten Zahlen sind Ereignisse, die unabhängig voneinander eintreten.

Wenn die Items lokal stochastisch unabhängig sind, kann das Ergebnis des Probanden in jeder ihrer Teilmengen eine erschöpfende Statistik für die ‚Tüchtigkeit‘ eines Probanden abgeben.

6.5.3 Stichprobenunabhängigkeit von Skala und Items (spezifische Objektivität, Teilgruppenkonstanz)

Wenn ein Itemsatz homogen ist, dann ist ein Test stichprobenunabhängig in einem besonderen Sinne. Genannt wird diese Eigenschaft auch „spezifische Objektivität“. Sie besagt:

Innerhalb der Population, für die Modellkonformität festgestellt ist, fallen für einen Probanden die Item- und Personenparameter immer gleich aus, gleichgültig welche Items er bearbeitet.

Gegenbeispiel: In einem Verfahren, das nach der klassischen Testtheorie konstruiert ist, können zwei Probanden ihre Rangplätze vertauschen, wenn man ihre Leistung nach Teilmengen von Items beurteilt.

Angenommen, Proband X habe 20 Items bearbeitet und dafür 15 Punkte erhalten, Proband Y habe dieselben 20 Items bearbeitet, aber nur 12 Punkte erreicht. Proband X nimmt Rangplatz 1, Proband Y Rangplatz 2 ein.

Aus der ‚Gesamtmenge‘ der 20 bearbeiteten Items greife man nun eine Teilmengen von 13 Items heraus und beurteile beide Probanden erneut:

- ⇒ Bei Proband X können jetzt 9 Antworten richtig und 4 falsch sein, so daß er nur 9 Punkte erhält,
 ⇒ Bei Proband Y können 12 richtig und 1 falsch sein.

In diesem Falle fiele Proband X auf Rangplatz 2, Proband Y rückte auf Rangplatz 1.

Ein solcher ‚Tausch‘ ist in einer Rasch-Skala nicht ‚zulässig‘: Items, die ihn zuließe, würden eliminiert.

Ein Itemsatz, der dem Raschmodell entspricht, ordnet *innerhalb der modellkonformen Population* die Probanden immer denselben Rangplätzen zu.

Gleichgültig, welche **Items** zur Messung ausgewählt werden, sie führen immer zu derselben Aussage über die Merkmalsausprägung derselben Probanden. „Hat eine Person im Gesamtest mehr Aufgaben gelöst als eine andere, dann kann sie in dem aus dem Gesamtest ausgewählten kleineren Itemsatz niemals weniger Items gelöst haben als diese andere Person (es kann nur sein, daß entweder die ursprüngliche Reihenfolge erhalten bleibt oder beide Personen jetzt die gleiche Zahl von Items gelöst haben)“ (Wottawa, 1980, 51-52).

Gleichgültig, welche **Personen** als Teilgruppe aus der modellkonformen Population ausgewählt werden, die Rangfolge unter ihnen ändert sich nicht.

Statt von ‚Stichprobenunabhängigkeit‘ spricht Wottawa auch *anschaulicher* von ‚Teilgruppenkonstanz‘ (1980, 53).

Einschränkung: Weil das Rasch-Modell einen probabilistischen Charakters hat, gelten die genannten Konsequenzen für die Stichprobenunabhängigkeit **im strengen Sinne** nur für den ‚Idealfall‘, d. h. nur für einen Itemsatz, bei dem gilt: Nie hat ein ‚Tüchtigerer‘ bei einem ‚leichteren‘ Item versagt, und nie hat ein ‚Unfähigerer‘ ein schwierigeres Item gelöst.

Im konkreten Falle eines empirischen Datensatzes, der ‚Rasch-skaliert wird‘, kann jedoch das Gegenteil eintreten: Probanden, die ‚tüchtiger‘ sind, können bei ‚leichteren‘ Items versagen und Probanden, die weniger ‚tüchtig‘ sind, können ‚schwerere‘ Items lösen; somit können auch zwei Probanden ihre Rangplätze tauschen.

6.5.4 Separierbarkeit von Item- und Personenparameter

In der klassischen Testtheorie sind Aussagen über Personen immer bezogen auf Items und ihre Lösungshäufigkeiten in einer Stichprobe. Alle Itemlösungen sind stichprobenabhängig. Ein Item hat keine ‚Schwierigkeit an sich‘, sondern immer nur eine ‚Schwierigkeit in einer bestimmten Personengruppe‘. Item- und Personenparameter treten in der klassischen Testtheorie zusammen auf, sie werden nicht getrennt konzipiert.

Im Rasch-Modell sind, wie in den Konstruktionsschritten veranschaulicht, die beiden Parameter als getrennte und als trennbare Größen konzipiert. Da die Items in einer modellkonformen Population bei allen Stichproben zu denselben Rangfolgen der Personen führen, hängt die Messung einer Tüchtigkeitsausprägung nicht von einer bestimmten Itemstichprobe ab.

Umgekehrt gilt dann: Die Parameter lassen sich trennen, indem in verschiedenen Personengruppen die gleichen Item- und Personenparameter ermittelt werden.

Man berechnet also die Parameter für verschiedene Stichproben unter der Vorannahme, daß innerhalb desselben Itemvektors der Itemparameter gleich und innerhalb derselben Probanden-Gruppe der Personenparameter gleich ist.

Nur solche Items und Personen werden als modellverträglich akzeptiert, für die diese Bedingungen zutreffen. ‚Unverträgliche‘ Items werden eliminiert, ‚unverträgliche‘ Probanden ausgeschieden.

6.6 Beitrag zu Diagnostik und Intervention

Mehr noch als für ‚klassische‘ Tests gilt für probabilistische Tests, daß sie sich mehr für Diagnostik eignen als für Intervention. Stichworte könnten lauten:

Beitrag zur Diagnostik: Probabilistische Tests

- ermöglichen eine wohldefinierte *Merkmalserfassung*,
- setzen im Idealfall ein Merkmal *in einen theoretischen Kontext*,
- ermöglichen eine *Vielzahl von Vergleichen*,
- tragen zur *Sprachregelung* zwischen den Diagnostikern bei.

Beitrag zur Intervention: Probabilistische Tests

- können helfen, *Interventionsbedarf* festzustellen,
- können die *Bilanzierung* einer Intervention erleichtern.

6.7 Kritische Anmerkungen

„Zweifellos ist das Rasch-Modell dem konkurrierenden klassischen Modell in bezug auf meßtheoretische Kriterien, die Möglichkeit zur empirischen Kontrolle der Modelleigenschaften und die Gewinnung erschöpfender Parameter überlegen“ (Michel & Conrad, 1982, 32). Rasch-Skalen wird das Niveau von Intervall-, gar von Rationalskalen zuerkannt.

Analog zum Modell der klassischen Testtheorie ist jedoch zu fragen, „ob Interpretationen der Parameterschätzungen angemessen sind, die über das **Ordinalskalenniveau** hinausgehen“ (Michel & Conrad, 1982, 32).

Dies gilt speziell für eine Interpretation von Item- und Personenparameter. Weil sie sich experimentell nicht trennen lassen, bleibt es schwierig, sie eindeutig zu ‚definieren‘ (Michel & Conrad, 1982, 32). Eine Interpretation, die dieselben ‚Gegebenheiten auslegt‘, aber ohne eine Annahme getrennter Item- und Personenparameter auskommt, müßte akzeptiert werden.

Ein schwieriges Problem stellt sich bei der Entscheidung, **Teilgruppen** zu bilden, an denen die Item- und Personenparameter gewonnen werden und die spezifische Objektivität sich bestätigen soll.

Wir haben in unserem Demonstrationsbeispiel die Gruppe A, B und C willkürlich gebildet. In der Praxis ist dies natürlich nicht möglich.

„Für die Wahl adäquater Teilungskriterien gibt es keine ‚Rezepte‘, so daß die Aussage daher auch immer nur heißen kann: ‚Modellkonformität‘ bei diesem oder jenem Teilungskriterium.“ Ein Erfolg mit einem bestimmten Kriterium garantiert nicht, daß „bei Wahl eines anderen Kriteriums... ebenfalls Modellkonformität resultiert“ (Guthke, Böttcher & Sprung, 1990, 165).

Darüber hinaus ist zu festzuhalten: Die **Art** der **Testkonstruktion** grenzt den Testgegenstand (die Eigenschaft, das Merkmal: den latenten trait) mit der Forderung nach der dem Rasch-Modell eigenen Homogenität sehr stark ein. Denn es treten **Selektionseffekte** auf, die sowohl die Items wie auch die Probanden betreffen: ‚Unangemessene‘ Items werden eliminiert, ‚unangepaßte‘ Probanden ausgeschieden, sie zählen nicht zu der Stichprobe, für die der Test modellkonform ist (Michel & Conrad, 1982, 32-33). „Aber ob es vom Begrifflichen her dann noch die Variable ist, die gemessen werden soll und ob diese Variable eine Vorhersageleistung für konkrete, in der Zukunft stattfindende Ereignisse bringen kann, geht aus dem Modell nicht hervor“ (Dieterich, 1973, 223).

Auf diese Weise kann sich die latente **Eigenschaft neu definieren**, ebenso die Stichprobe, für die der Test geeignet ist. Insofern kann es geschehen, daß ein Instrument für einen sehr schmalen Merkmalsausschnitt und eine sehr begrenzte Stichprobe sehr gute Meßergebnisse liefert, sich aber nicht generell anwenden läßt und keine diagnostisch relevanten Ergebnisse liefert.

Damit hängt die Frage zusammen, ob **Homogenität** im Sinne des Rasch-Modells die Homogenität im Sinne von Interkorrelation und Faktorenreinheit (im Sinne der Faktorenanalyse) einschließt. Das ist nicht der Fall. Insofern sind mit einer Rasch-Skalierung die Fragen nach Konsistenz, nach Reliabilität, Objektivität und Validität nicht erledigt.

Im Rasch-Modell ist nicht vorgesehen, die **‚klassischen‘ Gütekriterien** zu ermitteln. Die Informationen, welche die klassischen Gütekriterien geben, wer-

den von den Angaben des Rasch-Modells nicht ersetzt. (Dieterich, 1973, 223; Michel & Conrad, 1982, 32-34; Stelzl, 1972).

„Ebenso wie bei klassischen Tests sind zweifellos auch bei probabilistischen Tests, wenngleich unter anderen theoretischen Voraussetzungen, umfassende Validitätskontrollen unverzichtbar: Grundsätzlich ist die kontinuierliche Überprüfung empirischer Aussagen durch die Erfahrung ein Eckpfeiler einer jeden als Realwissenschaft verstandenen Disziplin... Unter diesem Blickwinkel sind Validitätsuntersuchungen auch im Rahmen probabilistischer Modelle notwendige Bewährungskontrollen. Solche Kontrollen sind aber auch deshalb angezeigt, weil die verfügbaren Itemselektionsalgorithmen und Verfahren der Modellkontrolle nicht ohne weiteres garantieren, daß die ermittelten Personenparameter streng valide Informationen über die ursprünglich in Frage stehenden latenten Dimensionen erbringen“ (Michel & Conrad, 1982, 54).

Die hohe Meßqualität genügt nicht, Rasch-Skalen als **„diagnostisch brauchbar“** zu erweisen. Denn bessere psychometrische Qualität verbürgt nicht schon höhere Adäquatheit für eine diagnostische Fragestellung. Neben der psychometrischen Qualität muß immer auch der diagnostische Nutzen (die Utilität) eines Verfahren berücksichtigt werden - der Beitrag, den ein Instrument zur Lösung einer Frage verspricht (Kap. 17, S. 373).

Ein neues Aufgabenfeld erschließen rasch-skalierte Tests, wenn sie für adaptives Testen eingesetzt werden. Darüber unterrichtet kurz Kapitel 18 (S. 391).

6.8 Zusammenfassung zu Kapitel 6

Das Rasch-Modell wurde entwickelt, um höheren meßtheoretischen Anforderungen zu genügen als die klassische Testtheorie.

Unterschieden werden Personenparameter und Itemparameter. Ihr Zusammenhang wird probabilistisch interpretiert: Dem Ereignis, daß ein Item gelöst wird, kommt eine bestimmte Wahrscheinlichkeit zu.

Ausgangsgleichung einer Skalierung ist die sogenannte Logistische Funktion, die den probabilistischen Zusammenhang zwischen Item- und Personenparameter abbilden soll.

Eine Rasch-Skalierung läßt sich in fünf Schritten zerlegen:

1. Es wird eine Matrix von Schwierigkeitsindizes erstellt.
2. Diese Matrix wird nach der Inversen der Logistischen Funktion in sogenannte Logits transformiert, welche die Differenz zwischen Personen- und Itemparameter repräsentieren.
3. Aus der Logit-Matrix werden die Personen- und Itemparameter geschätzt.

4. Anhand der geschätzten Personen- und Itemparameter wird eine Endmatrix der Schwierigkeitsindizes ermittelt, die sich auf Modellkonformität testen läßt.
5. Besteht die Endmatrix den Modelltest, dann ist es sinnvoll, die ermittelten Personen- und Itemparameter zu standardisieren.

Einer so konstruierten Rasch-Skala kommen bestimmte Charakteristika zu:

1. Homogenität der Items in einem spezifischen Sinn,
2. sogenannte lokale stochastische Unabhängigkeit der Items,
3. Stichprobenunabhängigkeit innerhalb der modellkonformen Population,
4. Separierbarkeit von Item- und Personenparameter.

Rasch-Skalen eignen sich - wie ‚klassische Tests‘ - mehr für diagnostische als für interventive Maßnahmen.

Kritische Anmerkungen betreffen Fragen wie die folgenden: Liegen die Testscores probabilistischer Skalen auf einem höheren Meßniveau als Ordinalwerte? Wird bei der Testkonstruktion nicht möglicherweise ein ‚neues‘ Testmerkmal konstruiert, wenn ‚unangemessene‘ Items eliminiert und ‚unangepaßte‘ Probanden ausgeschieden werden? Verbürgt die höhere psychometrische Qualität auch eine größere Adäquatheit für diagnostische Fragestellungen?

Ein neues Aufgabenfeld erschließen rasch-skalierte Tests, wenn sie für adaptives Testen eingesetzt werden.

6.9 Kontrollfragen zu Kapitel 6

- Grundannahmen.
- Ausgangsgleichung.
- Rolle der Schwierigkeitsindizes.
- Rolle der invertierten Ausgangsgleichung.
- Schritte einer Rasch-Skalierung.
- Trennung der Item- und Personenparameter.
- Bedeutung des Modelltests.
- Homogenität bei einer Rasch-Skala.
- Konzept der Stichprobenunabhängigkeit.
- Kritische Gesichtspunkte.

7. Kapitel

Verhaltensbeobachtung

Psychologie als empirische Wissenschaft definiert sich von der Verhaltensbeobachtung her. Insofern geht Verhaltensbeobachtung in jedes *psychologische* Handeln ein, auch in jeden *diagnostischen* Schritt.

Abgehoben von dieser Basisfunktion, erhält Beobachtung in Diagnostik und Intervention eine eigene Bedeutung. Sie kann *situative Bedingungen* einer Testung registrieren (etwa körperliche Symptome von Anstrengung oder Ermüdung). Sie kann Ergebnisse eines Leistungstests *ergänzen* (etwa, indem sie die Strategie erfaßt, nach der jemand Sortieraufgaben löst, oder dazu beiträgt, störende intervenierende Variablen zu identifizieren, die erwartungswidrige Testergebnisse „erklären“ könnten).

In diesem Kapitel sprechen wir über die zweite Konzeption von Beobachtung. Es geht um *Beobachtung als ein Verfahren neben anderen diagnostischen Verfahren*, nicht um Beobachtung als Basis aller psychologischen Methoden.

Den Stoff gliedern wir in sieben Teilkapitel:

- Abgrenzungen (Definitionen) (7.1),
- Festlegung von Beobachtungseinheiten (7.2),
- Einteilung der Verhaltensbeobachtung (7.3),
- Einfluß- und Verzerrungstendenzen (7.4),
- Beitrag zu Diagnostik und Intervention (7.5),
- Vor- und Nachteile der Verhaltensbeobachtung (7.6),
- zu den Gütekriterien der Verhaltensbeobachtung (7.7).

Das Kapitel schließt mit einer Zusammenfassung (7.8) und einer Reihe von Kontrollfragen (7.9).

7.1 Abgrenzungen (Definitionen)

Was heißt beobachten? Beobachten heißt, Ereignisse, Vorgänge oder Verhaltensweisen sorgfältig wahrnehmen und registrieren (Dorsch, 1994, 100; Faßnacht, 1995, 67-70; Huber, O., 1989, 124; Kaminski, 1977, 68-73; Selg & Bauer, 1971, 42). In Diagnostik und Intervention ist diese Wahrnehmung „un-

mittelbar und ausschließlich auf die psychologische Frage nach der Eigenart der individuellen Persönlichkeit gerichtet“ (Hasemann, 1983, 435).

„Unter wissenschaftlicher Beobachtung wird . . . die zielgerichtete und methodisch kontrollierte Wahrnehmung von konkreten Systemen, Ereignissen (zeitliche Änderungen in konkreten Systemen) oder Prozessen (Sequenzen von Ereignissen) verstanden“ (Huber O., 1989, 124).

In dieser Umschreibung heißt der Oberbegriff *Wahrnehmung*. Er deckt alle Arten sinnlicher Erfassung ab; Beobachten besagt demnach, menschliches Verhalten durch Sinneswahrnehmung zu erfassen: etwa durch Hören, Sehen, Tasten. In der Praxis dürfte die Verhaltensbeobachtung sich vorrangig beziehen auf eine Wahrnehmung mit Auge und Ohr, oft vermittelt durch spezielle Meßinstrumente.

Beobachtung bezeichnet eine besonders **aufmerksame Wahrnehmung**, die sich kontrolliert auf ihren Gegenstand richtet und das Ziel hat, eine genaue Kenntnis ihres ‚Gegenstandes‘ zu vermitteln.

Hauptgegenstand der Beobachtung sind in der Diagnostik

- erstens, Verhaltensweisen, die bei vielen Verfahren nicht eigens registriert werden, aber diagnostisch relevant erscheinen (z.B. Kommentare, die ein Proband zu den Items eines Fragebogens abgibt),
- zweitens, Verhaltensausschnitte, die eigens für die Beobachtung ausgewählt werden (z.B. eine Interaktionssequenz zwischen Mutter und Kind).

Selbst- und Fremdbeobachtung: Jede Beobachtung schließt Selbst- und Fremdbeobachtung ein.

Selbstbeobachtung (oder Introspektion) bezeichnet den Vorgang, in dem ein Beobachter sich selbst mit-wahrnimmt; die Selbstgegebenheit wird Mit-Gegenstand seiner Wahrnehmung; er ‚sieht‘ oder ‚hört‘ oder ‚fühlt‘ neben anderen Personen und Sachen *auch* sich selber.

Fremdbeobachtung bezeichnet die aufmerksame Wahrnehmung ‚anderer‘ Sachen oder Personen, die mit dem Beobachtenden nicht identisch sind.

Selbst- und Fremdbeobachtung sind einander komplementär zugeordnet: Wo keine Selbstbeobachtung, dort keine Fremdbeobachtung; wo Fremdbeobachtung, dort immer auch Selbstbeobachtung. Dies gilt in einem mehrfachen Sinne:

- Fremdbeobachtung schließt als mitlaufenden Prozeß das Mitbemerken des beobachtenden Subjektes ein.
- über den ‚Gegenstand‘ einer Fremdbeobachtung kann sich *ein einzelner* Beobachter nur dann mit einem *anderen* Beobachter verständigen, wenn der ‚Gegenstand‘ in seiner Selbsterfahrung ‚vorkommt‘ - vorkommt wenigstens in Erfahrungsspuren.
Einen ‚Gegenstand‘, welcher der Selbstbeobachtung eines Beobachters

völlig fremd wäre, könnte er nicht im Sinne einer Fremdbeobachtung ‚wahrnehme‘ - er könnte keine ‚Wahrnehmungsgestalt‘ bilden. (Ein Therapeut, der noch nie eine Spur von Angst verspürt hätte, wäre unfähig, bei seinem Klienten die Anzeichen von Angst wahrzunehmen und zu deuten.)

- Jede Beobachtung orientiert sich als Wahrnehmung an *Gesetzen*, die selber nicht auf Beobachtung gründen, weil jede Selbst- und Fremdbeobachtung sie schon voraussetzt.

Beobachtungen müssen sich in einen Begründungszusammenhang einfügen, sie müssen sich in Aussagen formulieren lassen, die widerspruchsfrei sind. Auf ‚Gesetzen‘ wie diesen beruht jede Kommunikation über Beobachtungen; diese Gesetze lassen sich jedoch nicht wieder aus der Beobachtung begründen - weil, um das Argument zu wiederholen, jede Selbst- und Fremdbeobachtung ihre Geltung schon voraussetzt.

Erschließbar sind solche Gesetze nur in jenem Erkenntnisvorgang, den wir Reflexion nennen. Beispiele sind erkenntnistheoretische Reflexionen, in denen sich ein Subjekt einem schon erkannten Objekt wieder und wieder zuwendet, um die Erkenntnistätigkeiten selber auszumachen - dabei allerdings immer auch auf Beobachtungsgegebenheiten angewiesen ist.

Was ist beobachtbar?

Was kann ein Untersucher beobachten? Was kann er *nicht* beobachten?

Beobachten kann er konkrete Verhaltensweisen: daß ein Proband redet, lächelt, aufsteht, umhergeht, sich niedersetzt, Fragen beantwortet, sich mit Aufgaben eines Tests beschäftigt...

Indes - kann der Untersucher wirklich *beobachten*, daß ein Proband sich mit Testaufgaben beschäftigt? Nein und Ja!

- **Warum Nein?** Weil der Untersucher nur wahrnehmen kann, daß der Proband vor dem Testheft sitzt und schreibt und... Nicht kann er ‚sehen‘, daß der Proband *sich beschäftigt* mit den Aufgaben. Es ist eine Schlußfolgerung, wenn der Untersucher ‚sieht‘, daß der Proband *sich* mit den Aufgaben *beschäftigt*. In jeder Beobachtung sind Interpretationen enthalten, Vor-Urteile (vorhergehende Urteile) - hier: die ‚Interpretation‘, daß ein Proband, der ein Testheft schriftlich bearbeitet, sich mit den Testaufgaben *beschäftigt*.
- **Warum Ja?** Solche Interpretationen gelten innerhalb bestimmter Referenzgruppen, hier der Psychologen, als Beobachtungen. Meist bleiben solche Abmachungen unausgesprochen. Explizit ausformuliert werden sie beispielsweise im Rahmen wissenschaftstheoretischer Reflexionen.

Nicht beobachten kann ein Untersucher, was einen Probanden befähigt, Testaufgaben zu lösen, oder was ihn bewegt, sich mit ihnen zu befassen. Fä-

higkeiten, Motive, Gefühle lassen sich nicht beobachten, sie lassen sich nur erschließen.

Dieser Schluß ist auf zwei Weisen möglich: Erstens, der Untersucher schließt aufgrund einer wohlbegründeten Regel, daß die Lösung einer bestimmten Anzahl von Aufgaben einen bestimmten Fähigkeitsgrad anzeigt, einen bestimmten Grad von Leistungsmotivation voraussetzt und von Gefühlen der Freude begleitet ist. - Zweitens, der Proband teilt dem Untersucher mit, was er kann, was er will und was er fühlt. - Jede der beiden Schlußweisen birgt ihre eigenen Schwierigkeiten.

Vorrang der Fremdbeobachtung: In der Psychologie ist mit Verhaltensbeobachtung vor allem Fremdbeobachtung gemeint. Dabei lautet ein Postulat, daß die Fremdbeobachtung **eines** (kompetenten) Beobachters kontrolliert werden soll durch die Fremdbeobachtung **anderer** (kompetenter) Beobachter.

Dieser Grundsatz betrifft vor allem die Forschung und dort solche Phänomene, die wiederholbar sind, oder solche, die mehrere Beobachter gleichzeitig wahrnehmen können. (Im strengen Sinne ist ‚gleichzeitige Wahrnehmung durch mehrere Beobachter‘ allerdings unmöglich, weil sich die Beobachtungspositionen verschiedener Beobachter nie vollständig decken: Bochenski, 1980, 64-65.)

Die Praxis (etwa einer Therapie) läßt eine Kontrolle derselben Phänomene durch mehrere Beobachter nur in begrenztem Maße zu. *Training* und *Supervision* übernehmen eine ähnliche Aufgabe wie die Kontrolle durch andere Beobachter.

Objektivität einer Beobachtung als Intentionalität: Wenn zwischen verschiedenen Beobachtern über denselben Gegenstand Konsens hergestellt wird, gilt eine Beobachtung als ‚objektiv‘. Ein solcher ‚intersubjektiver Konsens‘ kann eine Grundlage für weitere psychologische Aussagen abgeben.

Dieser Konsens ist ein intentionales Geschehen: „Nur als Objektivierungen intentionaler menschlicher Tätigkeit sind die Dinge und Ereignisse der Umwelt im strengen Sinne des Wortes humane Dinge und Ereignisse unserer Welt“ (Graumann, 1980, 49). Verhaltensbeobachtung ist zugeordnet einer psychologischen Umwelt, einem subjektiven Lebensraum - nicht einem physikalischen Umfeld (Lewin, 1963, 69; Thomae, 1968, 224).

Dieser Gedanke tritt noch klarer hervor, wenn wir die ‚Abgrenzung von Beobachtungseinheiten‘ besprechen. Solche Abgrenzungen setzen abstraktere kognitive Prozesse voraus als ‚bloßes Hinschauen oder Hinhören‘.

7.2 Festlegung von Beobachtungseinheiten

Im zeitlichen Ablauf eines Beobachtungsprozesses ist es eine der ersten Aufgaben zu bestimmen, „was beobachtet werden soll“, also eine Beobachtungseinheit festzulegen. (Die Frage sei auf die psychologische Diagnostik beschränkt.) Es bieten sich zwei Zugänge an: (1) deduktiver Weg, (2) induktiver Weg.

Zu (1): Beim deduktiv-theoretischen Zugang informiert sich der Untersucher, ob theoretische Konzepte oder gar Theorien vorliegen, die seine Abgrenzungen leiten können.

Angenommen, ein Untersucher wolle aggressives Verhalten beobachten, dann könnte er unterschiedliche Konzeptionen von Aggressivität sichten. Kasten 7-1 referiert Beispiele.

Kasten 7-1: Theoretische Konzeptionen von Aggressivität / Vier Beispiele

Freud definiert Aggression als Abkömmling und Hauptvertreter des Todestriebes, der im Organismus angelegt ist (Laplanche & Pontalis, 1977, 45).

Dollard et al. betrachten Aggression nicht als angelegten Trieb, sondern deuten sie als eine Reaktion. Dieser Grundansatz stellt sich in mehreren Varianten dar; an zwei Varianten sei erinnert: (a) Aggression ist **immer** Folge von Frustration. - (b) Frustration kann zu Aggression führen (1939).

Bandura erklärt Aggression als eine Verhaltensklasse, die an einem beobachteten Modell erlernt worden ist (1973).

Berkowitz, interpretiert Aggression als eine Verhaltensweise, die auf viele Auslöser zurückgeht, z. B. auf Ärger, Wut, Frustration. Zu dem Auslöser tritt immer ein kognitives Element hinzu, eine Interpretation. Beispielsweise kann der Anblick von Waffen aggressive Verhaltensweisen verstärken (1962).

Zu (2): Beim induktiv-empirischen Zugang sammelt der Untersucher ein Spektrum von Verhaltensweisen, die für die angezielte Beobachtungseinheit als relevant gelten - mit dem Ziel, sie ‚später‘ in einem theoretischen System zu ordnen.

„Diese Methode ist außerordentlich zeitraubend, sie läßt sich aber dann nicht vermeiden, wenn keine brauchbare Theorie zur Verfügung steht und man zunächst nach symptomatischen Verhaltensweisen suchen muß“ (Cranach & Frenz, 1969, 289).

Kasten 7-2 gibt *empirisch-induktiv* gesammelte Verhaltensweisen und Erklärungen wieder, die - gemäß bloßer Plausibilität - mit Aggression zu tun haben (könnten).

Kasten 7-2:
„Empirische“ Deutungen von Verhaltensweisen,
die mit Aggression zu tun haben (könnten)

- Aggressives Verhalten zeigt sich im Spiel: als spielerischer Kampf, der darauf abzielt, Stärke auszuprobieren oder Freude am Sieg zu erleben.
- Aggression stellt sich als Mittel dar, eigenes und fremdes Terrain zu erkunden.
- Aggressives Verhalten manifestiert sich als Abwehr von Bedrohungen.
- Aggressives Verhalten drückt sich aus in Handlungen, die sich „erklären“ als Rache für Niederlagen oder für Demütigungen.
- Aggressive Verhaltensweisen sollen Zuwendung, Aufmerksamkeit, Liebe „erzwingen“.

Deduktiver und induktiver Zugang schließen einander nicht aus, sondern können einander vorzüglich ergänzen.

Hilfen zur Abgrenzung von Beobachtungseinheiten

Wenn die Art des Zuganges (Deduktion vs Induktion) bestimmt ist: Wie gehe ich dann im einzelnen vor?

Was andernorts in der Diagnostik gilt, trifft auch hier zu: Es lassen sich „für die Auswahl sachlich angemessener Beobachtungseinheiten keine Kunstregeln vorgeben. Das Problem muß aus der konkreten Aufgabenstellung gelöst werden“ (Bungard, 1980, 80).

Doch lassen sich einige Hilfen benennen (Bungard, 1980, 78-81; Cranach & Frenz, 1969, 286-289; Faßnacht, 1995, 109-121; Hasemann, 1983, 461-463; Haubl & Spitznagel, 1983, 768-769; Huber, O., 1989, 124-143; Tismer, 1976, 832-833).

(1) Eindeutige Abgrenzungen suchen

Die Abgrenzung einer Einheit muß für die Beobachter eindeutig sein. Eindeutig sind Beispiele, die ein Merkmal markant kennzeichnen. Wie sich ein Konstrukt auf Eindeutigkeit prüfen läßt, dafür machen Schneewind und Dieterich Vorschläge (Kasten 7-3).

Kasten 7-3:
Zwei Vorschläge zur Einführung eindeutiger Begriffe

1. Schneewind (1969, 41) schlägt vor, Konstrukte anhand „repräsentativer Beispiele“ und „marginaler Gegenbeispiele“ einzuführen. „Repräsentative Beispiele“ sind Personen oder Objekte, die ein Merkmal in ausgezeichneter Weise repräsentieren: etwa ein Hypochonder für das Merkmal Hypochondrie. „Marginale Gegenbeispiele“ sind Personen und Objekte, die ein Merkmal nur peripher repräsentieren: etwa Depressive für das Merkmal Hypochondrie.

2. *Dieterich* (1973, 42-51) schlägt vor, eine Übungsgruppe von N Bewertern solle einer Beurteiltengruppe von k Probanden ein Konstrukt X zu- oder absprechen. Zu- oder Absprechen solle gekennzeichnet werden durch Eins oder Null. Die Werte lassen sich in eine Matrix eintragen, die es erlaubt, den Grad der Übereinstimmung numerisch zu bestimmen.

Rolle von Übung und Training: Eindeutige Abgrenzung setzt Übung und Training voraus. „Sorgfältige Beobachterschulung gehört in jedem Falle zur Entwicklung eines Beobachtungs-Instrumentariums bzw. zur Vorbereitung seines Einsatzes“ (Kaminski, 1977, 72).

Der Grad der Übereinstimmung zwischen Beobachtern ist eine Funktion

- einer *klaren Definition* der Beobachtungseinheiten,
- aber auch eines ausgiebigen *Trainings* der Beobachter.

Je öfter sich die Beobachter über wohldefinierte Beobachtungseinheiten verständigen, desto höher fällt ihr Konsens aus (Fisseni & Fennekels, 1995, 114).

Sprache und Beobachtung: Als spezielles Problem stellt sich hier der Zusammenhang von Sprache und Beobachtung (Faßnacht, 1979, 91-95; 1995, 105-108; Kaminski, 1977, 71-72; Selg & Bauer, 1971, 48).

Es geht um zwei Fragen:

1. Wie weit beeinflusst die Sprache unsere Beobachtungen? (Etwa - in welchem Maße *sehen* wir Sachverhalte, für die wir über *Wörter* verfügen?)
2. Wie angemessen bildet die Sprache unsere Beobachtungen ab? (Etwa - wie exakt können wir *sagen*, was wir *sehen*?)

Es liegt die Vermutung nahe: Wem mehr sprachliche Differenzierungen zu Gebote stehen, dem gelingen auch differenziertere ‚Verhaltensbeobachtungen‘ - im Doppelsinne des Wortes: Erstens, der Betreffende *nimmt* Sachverhalte differenzierter *wahr* Zweitens, er kann wahrgenommene Sachverhalte differenzierter *darstellen*.

Zwei Beispiele:

1. Die Eskimo-Sprache verfügt über etwa dreißig unterschiedliche Begriffe, um Zustände des Schnees zu differenzieren (Hartland, 1995, 228).
2. Die deutsche Sprache benennt das Phänomen der Liebe mit einem Wort, eben mit der Bezeichnung ‚Liebe‘. Die griechische Sprache führt drei Wörter an: Eros, Agápe, Philía. Das Indische verfügt über mehr als dreißig Bezeichnungen. - Die Kargheit oder die Vielfalt sprachlicher Unterscheidungen dürfte in beiden Fällen sowohl die ‚Sicht‘ der Sachverhalte als auch ihre ‚Darstellung‘ beeinflussen.

(2) Natürliche und künstliche Einheiten unterscheiden

Im Verhaltensstrom lassen sich Zusammenhänge ausmachen, die eine **natürliche Verhaltenseinheit** bilden: Murray spricht von *Episoden* (1938, 40), Thomaes von *Handlungen* (1968, 131-165). Episoden oder Handlungen sind Ver-

haltenssequenzen, bei denen sich klar ausmachen läßt, wann sie *beginnen*, wie lange sie *dauern* und wann sie *enden*.

Beispiel aus dem Umfeld der Aggression: Zwei Jungen tragen auf einem Spielplatz einen Ringkampf aus.

Andere Verhaltenseinheiten werden von außen festgelegt, es sind **künstliche Verhaltenseinheiten**. Prototyp ist das Experiment (Selg & Bauer, 1971, 54).

Als **Beispiel** seien, sehr gerafft, drei Phasen eines Aggressions-Experimentes referiert, das Bandura mit drei Gruppen von Kindern durchführte:

1. Probanden der Gruppe A sahen **einzeln** einen Erwachsenen, der eine große Puppe schlug, beschimpfte, mit Füßen trat. Probanden der Gruppe B sahen **einzeln** einen Erwachsenen, der ruhig im Experimentierraum bastelte, ohne die Puppe zu beachten. Gruppe C war die Kontrollgruppe, sie wurde weder der Bedingung A noch der Bedingung B ausgesetzt.
2. Alle drei Gruppen wurden in einen Raum mit Spielsachen geführt, aber schon nach zwei Minuten herausgerufen, sie wurden also frustriert, um auf diese Weise aggressiv gestimmt zu werden.
3. **Einzeln** wurden die Kinder dann in einen Nachbarraum geführt und sahen dort die große Puppe, aber auch anderes Spielzeug. Erwartet wurde, daß sich die frustrierten Kinder aggressiv verhalten würden; durch Einwegscheiben wurde ihr Verhalten beobachtet. Nun zeigte sich: Gruppe A (aggressives Modell) trat und schlug und beschimpfte die Puppe in weit höherem Maße als Gruppe C (Kontrollgruppe), erst recht als Gruppe B (friedliches Modell) (Bandura, Ross & Ross, 1961).

(3) Relevante Verhaltensweisen aussuchen

Es ist so gut wie unmöglich, alle Verhaltensweisen zu registrieren, die zu einer Beobachtungseinheit gehören (können), etwa das Gesamtrepertoire aggressiver Verhaltensweisen. Eine Auswahl ist unumgänglich, die Auswahl muß relevante Manifestationen betreffen.

Beispiele für Aggression: verbale Angriffe (Spöttelei, Schmähung, Ironie), ‚Handgreiflichkeiten‘ (Stoßen, Kneifen, Schlagen), Einsatz von Instrumenten (Stock, Riemen, Messer Pistole).

Was für eine konkrete Frage als zentral gilt, was als peripher, kann der Beobachter nur im Einzelfall entscheiden.

(4) Mit konkreten Beschreibungen beginnen, mit Abstraktionen abschließen

Es ist ein guter Ratschlag, an den Beginn einer Verhaltensbeobachtung konkrete Deskriptionen zu setzen, nicht abstrakte Interpretationen. Der Untersucher soll zunächst nur beobachten, noch nicht bewerten.

Beispiel:

1. Zunächst sollte ein Untersucher festhalten: „Während des Unterrichts stößt Stephan seinem Banknachbarn in die Seite oder kneift ihm in den Arm. Er stellt dem Lehrer Fragen, die nichts mit dem Unterrichtsstoff zu tun haben. Er schimpft leise vor sich hin.“
2. Erst aus einer Vielzahl solcher konkreter Angaben sollte das zusammenfassende Urteil abstrahiert werden: „Stephan verhält sich im Unterricht unkooperativ, er stört den Unterrichtsverlauf.“

Der Ratschlag, mit konkreten Angaben zu beginnen und mit Abstraktionen zu enden, beruht auf der Absicht, sich selbst und seine Mitbeobachter auf einen engen Interpretationsspielraum zu beschränken. Der Rat zielt darauf ab, die Schritte der Abstraktion zu kontrollieren.

(5) Bandbreite einer Beobachtungseinheit festlegen

Die Frage nach der Bandbreite ergänzt die Frage nach konkreten oder abstrakten Einheiten. Konkrete Verhaltensweisen sind eher mit schmalen Kategorien, abstrakte Eigenschaften eher mit breiten Kategorien erfassbar.

Beispiele:

1. Eher **schmale** Kategorien zu Aggression: brüllen, schimpfen, Klatsch verbreiten, eine andere Person schlagen, Tassen zu Boden schleudern, mit einem Stock Pflanzen köpfen.
2. Eher **breite** Kategorien: Spontane Aggressivität, Reaktive Aggressivität, Selbstaggression, Erregtheit bei Meinungsverschiedenheiten.

Zwei Gefahren: Definiert der Untersucher die Beobachtungseinheit zu schmal, kann er zwar viele Einzelereignisse registrieren; doch könnte es ihm schwerfallen, die Vielheit unter einem gemeinsamen Oberbegriff zusammenzufassen. - Definiert er seine Einheit zu breit, dann öffnet er den Freiraum für divergente Interpretationen.

(6) Disjunkte Einheiten anstreben

Beobachtungseinheiten sollen informativ sein. Am informativsten wären Kategorien, die über ein Merkmal alle Informationen enthalten. Aber solche Einheiten (wenn es sie denn gäbe!) brächten eine Schwierigkeit mit sich: Sie wären sehr komplex, müßten also ihrerseits in ‚kleinere‘ Einheiten aufgeglie-

dert werden. Unter den ‚kleineren Einheiten‘ wären dann solche Einheiten am hilfreichsten, die zu schon vorhandener Information ‚neue Information‘ hinzufügen, die also mit anderen Einheiten wenig Information teilen.

Mit anderen Worten, erstrebenswert sind Kategorien, die wenig Redundanz einschließen - vorteilhaft sind disjunkte Beobachtungseinheiten.

Disjunkte Einheiten stehen am Ende einer Untersuchung, nicht an ihrem Beginn - sie sind Ergebnis eines langen Filterprozesses. Am Beginn wird der Untersucher Überlappungen zulassen, erst allmählich die Redundanz seiner Einheiten vermindern.

Statistische Verfahren (z. B. Item-, Faktoren-, Clusteranalysen) können helfen, disjunkte Einheiten zu entdecken. Ob jedoch tatsächlich disjunkte Kategorien vorliegen, kann nur eine *logische* Analyse entscheiden.

(7) Vollständigkeit anstreben

In ein Beobachtungssystem sollten alle Verhaltensweisen eingehen, die als *relevant* gelten **für** die jeweilige Fragestellung. In diesem Sinne sollte das System umfassend und vollständig sein.

Beispiele:

1. Ein Autor beabsichtigt, einen Beobachtungsbogen zu konstruieren, mit dessen Hilfe er Aggressivität von Kindern erfassen will. Wenn er darauf abzielt, alle Arten von Aggressivität aufzunehmen, die bei Kindern vorkommen, muß er weit ausholen. Er darf keine gut belegten Phänomene ausklammern, die mit ‚kindlicher‘ Aggression zu tun haben, beispielsweise: Intention, etwas oder jemanden zu schädigen; offene Aggression (verbal, körperlich); verdeckte Aggression (phantasiert, sublimiert); kulturgebilligte Aggression (geschlechtsspezifisch orientiert), verurteilte Aggression (aus ethischen Gründen, aus sozialer Rücksicht oder bloßer Raffinesse). Und so weiter
2. Ein Untersucher ist bemüht herauszufinden, warum sich ein Vierzehnjähriger in der Schule höchst aggressiv aufführt, sich im Elternhaus jedoch höflich und beherrscht gibt. In diesem Falle braucht der Diagnostiker in seinem Untersuchungsplan nur solche Aggressionsformen vorzusehen, die - voraussichtlich - einen Erklärungswert haben; erkunden müßte er beispielsweise das soziale Umfeld des Jungen zuhause und in der Schule (Wird er anerkannt / gedemütigt?). Erfassen müßte er die kognitive Kapazität des Jungen (wird er überfordert / unterfordert?) Und so weiter

Das Ideal der Vollständigkeit begrenzt sich an der Fragestellung, die der Beobachter gewählt hat.

Abgrenzungsfrage - eine Einschränkung: „Es sei hier noch am Rande erwähnt, daß selbst eine klare Definition und enge Grenzen einer Beobachtungs-

einheit nicht notwendigerweise das tatsächliche Vorhandensein des damit umschriebenen Verhaltens voraussetzen. Genauso wenig ist damit impliziert, daß klar und eindeutig definierte Items auch die Möglichkeit zu ihrer unverfälschten Wahrnehmung bieten“ (Cranach & Frenz, 1969, 287).

7.3 Einteilung der Verhaltensbeobachtung

Die Verhaltensbeobachtung läßt sich unter vielfaltigen Gesichtspunkten einteilen. Besprochen seien folgende Klassifikationen:

- Systematische oder unsystematische Beobachtung (7.3.1),
- Beobachtung von Verlauf oder Zustand (7.3.2),
- Beobachtung in natürlicher oder künstlicher Situation (7.3.3),
- Teilnehmende oder nichtteilnehmende Beobachtung (7.3.4),
- Beobachtung von Zeit- oder Ereignisstichprobe (7.3.5),
- Beobachtung nach der Art ihrer Fixierung (7.3.6).

7.3.1 Systematische und unsystematische Beobachtung

Beobachtung nach ihrer Systematik einzuteilen heißt, den Übergang von vorwissenschaftlicher zu wissenschaftlicher Beobachtung zu betrachten. Dabei ist unsystematische Beobachtung nicht gleichzusetzen mit unwissenschaftlicher Beobachtung. Die Grenze ist nicht eindeutig zu ziehen.

Die **unsystematische Beobachtung** kommt der alltäglichen Beobachtung sehr nahe. Eine ‚Definition‘ ist deshalb schwierig. Soviel läßt sich sagen: Der Beobachtungsgegenstand ist (noch) nicht eindeutig festgelegt, (noch) nicht klar abgegrenzt. Der Beobachter hält seine ‚Eindrücke‘ eher selektiv fest; die Kodierung (soweit sie vorgenommen wird) ist eher deskriptiv (in Symbolen, Bildern, Worten), eher auch qualitativ (Abstufungen werden verbal angegeben, seltener in Meßwerten).

Beispiele: *Ein Vogelfreund hält tagebuchartig fest, was ihm am Verhalten der Vogel in seinem Garten auffällt. - Ein Ethnologe, am Anfang seines Aufenthaltes bei einem ‚Naturvolk‘, registriert, was er gerade sieht und hört, noch ohne eine Übersicht zu haben. - Ein Untersucher schaut einem Kind, das ihm vorgestellt wird, zunächst einfach zu (wobei das ‚einfache Schauen‘ allerdings strukturiert wird von seiner Erfahrung).*

Die **systematische Beobachtung** beruht darauf, daß Verlauf und Bereich der Beobachtung wohldefiniert sind. Festgelegt sind

- das Spektrum der Gegenstände, die beobachtet werden sollen,
- die Art der ‚Wahrnehmung‘ (mittels Auge allein, mittels Auge und Apparat, mittels Ohr und Tonband),

- die Art der Fixierung (auf Film, auf Tonband, auf Papier),
- die Prozedur der Auswertung (ausgearbeitetes Kategoriensystem).

Das perfekte Paradigma der systematischen Beobachtung ist das psychologische Experiment.

Beispiel (bei dem es um ein diagnostisches Experiment geht oder um eine dem Experiment ähnliche Anordnung):

Ein Rehabilitand sitzt am „Wiener Determinationsgerät“, das ihm zweierlei vorgibt: erstens fünf Kreise, die optische Reize bieten können (Weiß, Gelb, Rot, Grün, Blau), zweitens Tasten in den Farben der fünf Reize. - In zufälliger Reihe oder in systematisch variiert Sequenz erscheinen diese farbigen Reize. Aufgabe des Probanden ist es, jeweils jene Taste zu drücken, deren Farbe dem Reiz entspricht. - Beobachtet werden Tempo und Exaktheit der sensu-motorischen Reaktion.

Zwischen systematischer und unsystematischer Beobachtung sind viele **Mischformen „halb-systematischer Beobachtung“** denkbar. Ihre Ausgestaltung orientiert sich am konkreten Ziel: Ein Teil der Beobachtungen, ihrer Registrierung, ihrer Auswertung ist fixiert, ein Teil frei variierbar.

Beispiel: *Bei der Untersuchung aggressiver Verhaltensweisen eines Sechsjährigen kann der Diagnostiker standardisierte Situationen einplanen, die Aggressivität wecken sollen (Wegnahme von Spielzeug, wenn ‚das Spiel am schönsten‘ ist). Zusätzlich kann er in diesen Sequenzen ‚freie Beobachtungen‘ vorsehen. - Darüber hinaus kann er Aggressivität in unstandardisierten Situationen beobachten, etwa wenn der Junge malt oder wenn er Spielzeug bekommt oder mit Figuren und Objekten aus dem Steno-Testkasten spielt. - Zusätzlich kann er für diese offeneren Situationen auch Sequenzen festlegen, die standardisiert beobachtet und kodiert werden. - Auf diese Weise sind vielfältige Kombinationen möglich.*

Andere Terminologie: Statt von systematischer und unsystematischer kann man auch sprechen

- von *standardisierter* und *unstandardisierter*;
- von *strukturierter* und *unstrukturierter* oder
- von *kontrollierter* und *unkontrollierter* Beobachtung.

Anwendung: Dem diagnostischen Prozeß kommt Beobachtung in allen ihren Varianten zustatten:

- Zur Einzelfalldiagnostik liefern sowohl die systematische oder unsystematische wie auch die halb-systematische Beobachtung ihren Beitrag.
- Für Forschung und Reihenuntersuchung empfiehlt sich vor allem die systematische Verhaltensbeobachtung.

Kasten 7-4 veranschaulicht an einem Beispiel, wie ein Beobachter aus einer unsystematischen Beobachtung eine systematische entwickelt.

Kasten 7-4:**Beispiel für den Übergang von unsystematischer zu systematischer Beobachtung**

Ein Landwirt, der ökologischen Landbau betreibt, berichtet:

„Vor vielen Jahren machte ich . . . folgende Beobachtung: in einer Pfirsichanlage von etwa einem Morgen Größe blieben einzelne Bäume immer frei von Läusen, selbst wenn ringsum stehende Bäume davon befallen waren. Was mochte der Grund sein? Bei näherer Betrachtung stellte sich heraus, daß der einzige feststellbare Unterschied der war, daß diese Bäume vom Rainfarn wachsen waren, der ansonsten in meiner Plantage kaum vorkam. Da der Standort der gleiche war und auch die Bäume sonst gleich behandelt worden waren, mußte dies der Grund für die besondere Widerstandsfähigkeit gegen Läusebefall sein. Sofort pflanzte ich Rainfarn auch an die anderen Pfirsichbäume, und der Erfolg trat nach zwei Jahren ein“ (Erven, 1981, 79).

7.3.2 Beobachtung von Verhaltensverlauf oder Verhaltenszustand

Die Verhaltenseinheit, die beobachtet werden soll, kann sich auf einen **Verlauf** oder einen **Zustand** beziehen.

Verlauf: Gegenstand der Beobachtung kann die Art des Vorgehens bei einer Problemaufgabe sein, etwa bei der Bearbeitung des Mosaiktests (HAWIE) oder die Art der Auswahl von Klötzchen bei der ‚Sortierprobe nach Stein‘. Der Verlauf selber soll Aufschluß geben über Verhaltensstrategien. - Besonders bei Partnerschaftskonflikten trägt es zum Verständnis bei, die Entstehung des Konfliktes zu erfassen; allerdings durfte in der diagnostischen Situation der Konflikt-Verlauf meist nur *rekonstruierbar* sein (Abruf von Erinnerungen); dies wirft eigene Probleme auf.

Zustand: Es soll Verhalten erfaßt werden, das sich zusammenfassen läßt als Fertigkeit oder Fähigkeit, etwa unter Titeln wie Konzentration oder Intelligenz. In einer Eignungsuntersuchung könnte der Untersucher darauf abheben, den maximalen Grad einer Fähigkeit zu erfassen.

Anwendung: Das Interesse des Untersuchers kann wechseln. Einmal kann es stärker am beobachtungsnahen Ablauf haften, sich ein andermal stärker zum abstrahierten Wesenszug hin verlagern.

7.3.3 Beobachtung in natürlicher oder künstlicher Situation

Die Umstände einer Verhaltensbeobachtung lassen eine Unterscheidung von natürlicher und künstlicher Situation zu:

1. Beobachtung geschieht **in einer natürlichen Situation**, wenn der Untersucher die Situation nimmt, wie sie vorgegeben ist - er manipuliert sie nicht. (Das Gegenteil ist die Laborsituation, das Experiment.) Die Beobachtung in natürlicher Situation ist zweifach möglich:

- Sie kann unmittelbar gegeben sein, indem Erleben und Beobachtung parallel laufen. **Beispiel:** Ein Untersucher, der einen Sorgerechtsfall

bearbeitet, beobachtet die zwölfjährige Tochter des betreffenden Paares in der Schule (ohne Wissen des Kindes).

- über Verhaltensbeobachtungen kann aus der Retrospektive berichtet werden. Solche Beschreibungen unterliegen immer der Gefahr einer Informationsverzerrung. **Beispiel:** Eltern geben Auskunft über die Entwicklung ihres Kindes. Probanden berichten über frühere Abschnitte ihres Lebens.

2. Beobachtung findet **in künstlichen Situationen** statt, wenn der Beobachter wesentliche Bedingungen des Verhaltens kontrollieren kann. Idealtyp ist das psychologische Experiment.

Sonderfall - die ‚gestellte Situation‘: Ein Sonderfall der Beobachtung in künstlicher Situation ist die Verhaltensbeobachtung in gestellter Situation. Sie ist gegeben, wenn der Proband eine Situation als natürlich erlebt, der Untersucher sie aber zum Zwecke der Verhaltensbeobachtung ‚konstruiert‘ und in diesem Sinne ‚gestellt‘ hat.

Beispiel: Ein Untersucher sieht eine Situation vor, in der ein Proband, der sich um eine Stelle bewirbt, im Wartezimmer mit einer Person spricht, die sich als Mitbewerber ausgibt, in Wirklichkeit jedoch ein Mitarbeiter des Untersuchers ist.

Anwendung: Wo in der diagnostischen Situation spielen die drei Arten eine Rolle?

- Beobachtung in natürlichen Situationen dürfte bei vielen Untersuchungsanlässen empfehlenswert sein.
- Beobachtung in (künstlichen, in) kontrollierten Situationen dürften im diagnostischen Prozeß die Regel sein.
- Beobachtung in gestellten Situationen sind im diagnostischen Prozeß zwar denkbar, dürften aber die Ausnahme bleiben.

7.3.4 Teilnehmende und nicht-teilnehmende Beobachtung

Nach dem Grad, in dem sich der Untersucher selber in das beobachtete Geschehen einbezieht, lassen sich teilnehmende und nicht-teilnehmende Beobachtung unterscheiden.

Als **teilnehmende Beobachtung** gilt der Fall, in dem der Beobachter selber Teil des beobachteten Geschehens ist. Paradigma dafür ist der Anthropologe, der bei ‚Naturvölkern‘ als Gast lebt und als solcher seine Aufzeichnungen macht. Ein anderes Beispiel ist die Beobachtung der „Norton Street Gang“ durch Whyte, der sich als Mitglied in diese „Street Corner Society“ hatte einführen lassen (Whyte, 1943).

Nicht-teilnehmende Beobachtung liegt vor, wenn der Beobachter sich vom Geschehensablauf abhebt. Ein extremes Beispiel ist die Beobachtung von Probanden durch ein Einweg-Fenster.

Anwendung: Im diagnostischen Prozeß dürfte die nicht-teilnehmende Beobachtung die Regel sein, die teilnehmende dagegen die Ausnahme.

7.3.5 Erfassung von Zeit- oder Ereignisstichprobe

Eine weitere Einteilung ergibt sich aus der Frage, welcher Verhaltensausschnitt beobachtet werden soll.

Von **Zeitstichprobe** spricht man, wenn in festgelegten Zeitabschnitten ‚alles Verhalten‘ beobachtet wird (time-sampling).

Beispiel: Filmaufnahmen werden gemacht

1. während der gesamten Untersuchung mit dem Ziel, das Gesamtverhalten zu erfassen,
2. nur in einigen festgelegten Phasen einer Eltern-Kind-Interaktion, beispielsweise alle fünf Minuten dreißig Sekunden lang.

Ereignisstichproben werden registriert, wenn spezifische Verhaltensausschnitte abgegrenzt und nur sie beobachtet werden - andere Verhaltensanteile also ausgeblendet bleiben (event-sampling).

Beispiele:

1. Bei der Aufgabe, gemeinsam zu telefonieren, wird von den Eltern-Kind-Interaktionen nur ‚Belohnung‘ oder ‚Bestrafung‘ registriert.
2. Ein Klient erhält die Aufgabe zu registrieren, was alles sich um ihn herum abspielt, wenn seine Stimmung sinkt und depressive Züge annimmt.

Anwendung: Im diagnostischen Prozeß dürfte während der Untersuchung die Ereignisstichprobe überwiegen. - Ein Untersucher kann auf eine Zeitstichprobe abheben, wenn er erfahren will, ‚was alles‘ in einem bestimmten Lebensabschnitt geschehen ist.

7.3.6 Verhaltensbeobachtung nach der Art ihrer Fixierung

Wichtig, vor allem für eine Auswertung, ist die Art, in der Beobachtungen festgehalten werden.

Eine **mechanische Registrierung** ist möglich auf Film, auf Tonband oder mithilfe anderer Apparate.

Eine **symbolische Fixierung** ist auf verschiedene Weise möglich:

- Beobachtungen werden in einer freien Beschreibung formuliert (anekdotische Beschreibung: Hasemann, 1983, 443).

- Besondere *Ratingskalen* werden konstruiert, einzelne Punkte mit numerischen (gegebenenfalls auch mit verbalen) Ankern versehen (Hasemann, 1983, 445-446, 453-463).
- *Andere Zeichensysteme* werden entworfen, in ihnen Beobachtungen fixiert (Cranach & Frenz, 1969, 310-3 11).
- Es werden *standardisierte Beobachtungsbögen* übernommen (Hasemann, 1983, 446-448) beispielsweise
 - ⇒ der „Diagnostische Elternfragebogen“ (DEF) von Dehmelt, Kuhnert & Zinn (1981) oder
 - ⇒ der „Beobachtungsbogen für Kinder im Vorschulalter“ (BBK) von Duhm & Althaus (1979) oder
 - ⇒ das bekannte Interaktionsschema von Bales (Interaktions-Prozeß-Analyse: 1950).

Anwendung: Bei der Anwendung gilt es, zu unterscheiden:

- In der *Individualdiagnostik* (vermutlich auch bei Reihenuntersuchungen) dürfte - vor allem aus Zeitgründen - die sprachliche Fixierung überwiegen.
- In der *Forschung* sollte sich die mechanische Registrierung durchsetzen, etwa als Tonbandmitschnitt, im Idealfall als Videoaufzeichnung.

Kasten 7-5 bringt Beispiele aus Beobachtungsbögen für Kinder.

**Kasten 7-5:
Beobachtungsbögen - kurze Auszüge**

1. Beispiel:

Aus der „Hamburger Verhaltensbeurteilungsliste (HAVEL)“ von Wagner (1981):

1. Das Kind spielt lieber drinnen als draußen.
2. Das Kind neigt zu Wutausbrüchen.
3. Das Kind leidet an Übelkeit oder Erbrechen.
4. Das Kind behandelt seine Schulbücher sorgfältig.
5. Wenn das Kind etwas tun soll, fällt es ihm schwer, den Anfang zu finden.

Auswertung: Vorgegeben werden fünf verschieden große Kreise, die anzeigen, wie häufig eine Aussage auf das Kind zutrifft. (Größter Kreis: „Immer oder fast immer.“ - Kleinster Kreis: „Nie oder fast nie.“)

2. Beispiel:

Aus dem „Arzt-Kind-Interaktionsbogen“ von Petermann (1985, 83): „Kategoriensystem Kind“:

A. Verbales Verhalten:

- (a) Fragen stellen vs. keine Fragen stellen.
- (b) Fragen des Arztes beantworten vs. nicht antworten.
- (c) Erzählen eigener Erlebnisse vs. kein selbständiges Erzählen.
- (d) Gefühlsäußerungen vs. keine Gefühle zeigen.

B. Nonverbales Verhalten:

- (e) Blickkontakt vs. kein Blickkontakt.
- (f) Sich ohne Scheu anfassen (untersuchen) lassen
vs. sich nicht anfassen lassen.
- (g) Still sein mit oder ohne Blickkontakt.

C. Tätigkeiten:

- (h) Sich umsehen vs. sich nicht umsehen.
- (i) Ausprobieren von Instrumenten vs. kein Ausprobieren.
- (k) Den Anweisungen des Arztes folgen vs. nicht folgen.

D. Sonstiges Verhalten:

- (l) Restkategorie.

Auswertung: Es wird „nur erhoben, ob ein Merkmal vorliegt oder nicht“ (1985, 82).

Kommentar zu Kasten 7-5:

1. Solche Bögen bieten dem Untersucher Hilfen für seine Beobachtungen - auch zur eigenständigen ‚Definition‘ von Beobachtungseinheiten.
2. Doch sollte er sich klarmachen: Jede der vorgeschlagenen ‚Beobachtungen‘ beruht auf Vor-Einstellungen und Abgrenzungen der Autoren.
3. So bleibt immer zu fragen:
 - ⇒ Welche Beispielfragen beziehen sich auf einzelne sichtbare oder hörbare Verhaltensweisen?
 - ⇒ Welche Beispielfragen enthalten Zusammenfassungen, also Interpretationen über mehrere Situationen hinweg?

7.4 Einfluß- und Verzerrungstendenzen

Verhaltensbeobachtung wurde als Wahrnehmung ‚definiert‘. Wahrnehmungsprozesse laufen selektiv ab: Aus dem Wahrnehmungsfeld schneidet die wahrnehmende Person Segmente aus und organisiert sie zu einer ‚Gestalt‘. Diese Organisation wird vielfältig beeinflusst.

Einige Einflußgrößen seien aufgezählt, sie heben sich nicht disjunkt voneinander ab, sondern überschneiden sich (Anger, 1969, 598; Bungard, 1980, 22-26; Cranach & Frenz, 1969, 274, 280-285; Faßnacht, 1995, 220-229; Hase-mann, 1983, 463-472; Huber, O., 1989, 137-138; Maccoby & Maccoby, 1965, 74-76; Schram & 1964, 888; Tismer, 1976, 829-831; Wottawa & Hossiep, 1986, 86).

Wir unterscheiden zwei Klassen

- Fehler allgemein bei diagnostischen Verfahren (7.4.1),
- Fehler speziell bei der Verhaltensbeobachtung (7.4.2).

7.4.1 Allgemeine Fehler

Mit allgemeinen Fehlern bezeichnen wir Verzerrungstendenzen, die bei den meisten diagnostischen Verfahren auftreten können: bei Verhaltensbeobachtungen ebenso wie bei Gesprächen, bei der Vorlage eines Fragebogens ebenso wie bei der Durchführung eines projektiven Verfahrens, ja sogar bei der Bearbeitung von Leistungstests.

Wir zählen sechs Beispiele auf.

Hofeffekt oder Überstrahlungseffekt: Eine zentrale Eigenschaftsdimension bestimmt die Eindrucksbildung. Dieser Einfluß manifestiert sich darin, daß unter mehreren Eigenschaften *eine* Eigenschaft so dominiert, daß sie die Anordnung der anderen (mit-) bestimmt.

Beispiel: Exemplarisch lassen sich Experimente von Asch anführen (1946, 262-263). Zwei Gruppen A und B erhielten Eigenschaftslisten. Mit den Eigenschaften sollten sie eine Person beschreiben. - Die Liste der Gruppe A lautete: *Intelligent, skilful, industrious, warm, determined, practcial, cautious.* - Die Liste der Gruppe B lautete: *Intelligent, skilful, industrious, cold, determined, practcial, cautious.* - Die beiden Listen unterschieden sich also nur in einem Wort, bei A stand ‚warm‘, bei B dagegen ‚cold‘.

Ergebnis:

- Bei Gruppe A zentrierten sich die Wörter um den Pol ‚warm‘. So wurde eine Person X beschrieben wie folgt: „A person who believes certain things to be right, wants others to see his point, would be sincere in an argument and would like to see his point won“ (1946, 263).
- Umgekehrt bei Gruppe B: Die Beschreibungen zentrierten sich um das Adjektiv ‚kühl‘. Eine Person Y wurde charakterisiert wie folgt: „A rather snobbish person, who feels that his success and intelligence set him apart from the run-of-the-mill individual. Calculating and unsympathetic“ (1946, 263).

Positions-Effekt: Erster oder letzter Eindruck steuert die gesamte Beurteilung. Eine erste (oder eine letzte) Information prägt das Urteil.

Beispiel: Eine Studentin erscheint elegant gekleidet zur Prüfung, sie redet flüssig und selbstsicher. Der Prüfer **beeindruckt**, attribuiert ihr von vornherein eine hohe Intelligenz.

Milde-Effekt, Strenge-Effekt: Der Fehler, der seinen Namen von der ‚Milde‘ erhält, wird unterschiedlich beschrieben.

1. Angespielt wird auf die Tendenz, generell günstige Urteile abzugeben (generosity error).
2. Gemeint ist die Neigung, solche Personen günstig zu beurteilen, die dem Beurteiler sympathisch sind (leniency error). - Wird sich ein Beurteiler seiner Milde-Tendenz bewußt, könnte er zum Gegenteil neigen, zu überstrengen Urteilen.

Beispiele:

Zu 1.: Ein Lehrer gibt generell ‚gute‘ Noten (generosity).

Zu 2.: Bei Schülern, die einem Lehrer sympathisch sind, konvergieren die Noten bei Gut (leniency).

Zentrale Tendenz: Der Beurteiler bevorzugt neutrale und meidet extreme Urteile.

Beispiel: Bei Einstufungsmethoden (etwa Ratings) zentrieren sich die Scores im Mittelbereich.

Kontrastfehler, Ähnlichkeitsfehler:

1. Der Befragter neigt dazu, beim Befragten Eigenschaften zu ‚erkennen‘, die er sich selber abspricht (Kontrastfehler).
2. Der Befragter könnte aber auch dazu neigen, beim Befragten Eigenschaften zu ‚entdecken‘, die er sich selber zuschreibt (Ähnlichkeitsfehler).

Beispiel:

Zu 1.: Ein Untersucher, der bei sich selber ‚weiches Verhalten‘ unterdrückt, könnte dafür sensibilisiert sein, bei einem Gesprächspartner weiche Verhaltensweisen ‚festzustellen‘ (Kontrastfehler).

Zu 2.: Ein Untersucher der sich für einfühlsam hält, könnte dazu tendieren, bei einem Gesprächspartner Empathie zu identifizieren - ohne weitere Überprüfung (Ähnlichkeitsfehler).

Erwartungs-Effekt (selffulfilling prophecy): Ein Beurteiler läßt sich in seinen Schlußfolgerungen von ungeprüften Hypothesen leiten. „Dieses Verhalten ist der Neigung zum ‚Stereotypisieren‘ verwandt“ (Tismer, 1976, 830).

Beispiel: Ein Untersucher sei der Meinung, der zweite Bildungsweg vermittele eine undifferenzierte Schulbildung. Nun untersuche er einen Probanden, der den zweiten Bildungsweg gegangen ist. Aufgrund seiner Vorannahme könnte der Untersucher dazu neigen, den Probanden als undifferenziert zu klassifizieren - ohne weitere Überprüfung.

7.4.2 Fehler speziell bei der Verhaltensbeobachtung

Mit speziellen Fehlern bezeichnen wir Verzerrungstendenzen, die sich im besonderen auf die Verhaltensbeobachtung beziehen - leider jedoch nicht ausschließlich auf sie (Huber, O., 1989, 137).

Überforderte Differenzierungsfähigkeit: Die menschliche Wahrnehmungskapazität ist begrenzt - wie die Allgemeine Psychologie belegt. Ein Untersucher überfordert seine Kapazität, wenn er sich vornimmt, ‚zu viele‘ Objekte zu beobachten.

Beispiel: Ein Untersucher legt zu viele Kategorien fest, unter denen er einen Probanden beobachten möchte. - Oder er wählt eine zu große Anzahl von Probanden, denen er seine „besondere Aufmerksamkeit“ zuwenden will.

Unschärfe Definition: Beobachtungseinheiten, die unscharf umrissen sind, gewähren zuviel Freiraum für Interpretationen. Der Beobachter bestimmt eigenständig, aber auch eigenmächtig, was er ‚sieht oder hört‘.

Beispiel: Ein Untersucher erhält den Auftrag, Leistungsmotivation zu beobachten. Ein Gegenstand, der ‚Leistungsmotivation‘ heißt, ist für eine Beob-

achtung zu breit und zu unscharf gefaßt. Das Merkmal müßte in kleinere Einzelfacetten aufgegliedert werden.

Unvertrautheit mit den Beobachtungseinheiten: Ein Untersucher beherrscht die Beobachtungseinheiten nicht oder ist nicht vertraut mit dem Kode der Aufzeichnung.

Beispiel: *Ein Beobachter in einem Assessment-Center ist nicht hinreichend geschult (worden), das Anforderungsprofil differenziert zu interpretieren und anzuwenden.*

Unvertrautheit mit der Probanden-Gruppe: Der Untersucher ist unvertraut mit den Normen und Standards der Probandengruppe, die er beobachten soll.

Beispiel: *Ein Gutachter beurteilt einen Strafgefangenen, ohne das Milieu des ‚Knasts‘ zu kennen.*

Eingriff in den Untersuchungsablauf: Das Verhalten des Beobachters weicht ab von dem Verhalten, das für ihn festgelegt worden ist.

Beispiel: *Ein Untersucher befolgt nicht die vorgegebene Instruktion; während einer Untersuchung, beispielsweise bei einem Test, gibt er Rückmeldungen über falsche oder richtige Aufgabenlösungen - entgegen der Instruktion.*

Resümee: Die verschiedenen Fehlertendenzen lassen sich nicht disjunkt trennen. Der gemeinsame Zug dürfte in der Neigung liegen, über Personen Urteile abzugeben, deren Richtigkeit nur angenommen, aber nicht festgestellt wird.

„Die Beobachtung einer Person unterliegt der . . . allgemeinen Neigung des Menschen, andere Menschen aufgrund einer ihm eigenen Meinung von vorneherein mit einer gewissen individuellen Tendenz zu beurteilen (implizite Persönlichkeitstheorie).“ Dabei gilt, „daß Beobachter um so weniger zwischen Personen differenzieren können, je einfacher ihre implizite Persönlichkeitstheorie ist“ (Hasemann, 1983, 464).

Die **Folgerung**, die zu ziehen ist, liegt nahe: „Die wirkungsvollste Methode zur Vermeidung von systematischen Fehlern und zur allgemeinen Verbesserung der Beurteilungen, ist wohl eine sorgfältige Schulung der Beurteiler“ (Hasemann, 1983, 471).

Eine Chance, Fehler gleichsam erfahrbar zu machen und ein Training als notwendig erleben zu lassen, bieten Ton- und Video-Aufzeichnungen:

- Aufzeichnungen können Beobachtungsfehler demonstrieren.
- Aufzeichnungen können dazu beitragen, Fehler ‚am konkreten Fall‘ zu korrigieren.

7.5 Beitrag zu Diagnostik und Intervention

Sowohl für die Diagnostik als auch für die Intervention kann die Verhaltensbeobachtung einen speziellen Beitrag liefern.

Was leistet die Beobachtung für die Diagnostik?

In der Diagnostik kann die Beobachtung auf zweierlei Art zur Verhaltens erfassung beitragen:

- als begleitendes und
- als selbständiges Verfahren.

Als **begleitendes Verfahren** läßt sich Verhaltensbeobachtung dazu verwenden, den situativen Kontext anderer Verfahren festzuhalten. Dabei kann sie Informationen liefern, die es erlauben, die Ergebnisse anderer Verfahren zu evaluieren.

Beispiele:

1. *Bei einer Anamnese kann ein Widerspruch auftreten zwischen zwei Verhaltensebenen - etwa zwischen verbalen Äußerungen und manifestem Verhalten. So kann der Inhalt von Aussagen, die Eltern über ihre Kinder machen, in Widerspruch stehen zu der Art, in der sie ihre Aussagen vorbringen. Wird allein der Inhalt des Gespräches registriert, dann bleibt dieser Widerspruch unberücksichtigt, obwohl gerade er diagnostisch wichtige Signale gibt.*
2. *Bei manchem Test erlaubt die Art des Vorgehens Schlüsse auf Problemlösestrategien des Probanden. So kann es aufschlußreich sein, zu beobachten, wie ein Kind im Mosaiktest des HAWIK die Muster legt, probierend oder planend, systematisch oder unsystematisch.*
3. *Bei der Beantwortung eines Fragebogens kann es entscheidend sein, die Kommentare des Probanden zu registrieren. In ihnen kann sich Ablehnung ebenso wie Zustimmung, aber auch Verzerrung zu erkennen geben. Wenn beim ‚Gruppentest für die soziale Einstellung‘ (SET: Joerger, 1973) ein Zehnjähriger erklärt: „Ich weiß, daß Antwort b) auf mich mehr zutrifft, aber Antwort c) hören Sie lieber“, dann entwertet ein solcher Kommentar den Test-Score, er ist nicht verwendbar.*

Resümee: Verhaltensbeobachtung als begleitendes Verfahren kann mitentscheiden, ob Ergebnisse anderer Verfahren ‚valide im Einzelfalle‘ sind.

Als **selbständiges Verfahren** läßt sich Verhaltensbeobachtung vielfältig in die diagnostische Untersuchung eingliedern:

Beispiele:

1. *Die Körpersprache von Ehepartnern kann zum Gegenstand gezielter Beobachtung gemacht werden. Günstig ist es, wenn zwei Untersucher beteiligt sind, von denen der eine das Gespräch führt, der andere das Verhalten beobachtet.*

2. *Ebenso lässt sich die Eltern-Kind-Interaktion zum Thema machen, etwa bei gemeinsamen Rollenspielen oder bei der Vorlage des Steno-Tests (Staabs, 1964).*
3. *Rollenspiele können einen eigenen Part in der diagnostischen Situation erhalten, bis hin zu dem Versuch, eine Anforderungssituation im Rollenspiel möglichst getreu abzubilden.*

Resümee: Verhaltensbeobachtung als selbständiges Verfahren kann, besonders in der Individualdiagnostik, andere Verfahren wesentlich ergänzen.

Was leistet die Beobachtung für die Intervention?

In der Intervention kann die Beobachtung ‚Eingriffs-Stellen‘ identifizieren, kann dazu beitragen, die Art eines Eingriffs zu definieren, schließlich kann sie das Ergebnis von Eingriffen dokumentieren.

- *Ein Beispiel aus der **Arbeitspsychologie**: Im Assessment-Center dient die Beobachtung als wichtigste, im Idealfall sogar als einzige Erfassungsmethode. Erfasst werden soll beispielsweise das Fähigkeits- und Entwicklungspotential eines Teilnehmers. - Aufgrund von Verhaltensbeobachtungen werden „Empfehlungen“ formuliert, deren Befolgung dazu beitragen soll, das gegebene Potential weiterzuentwickeln (Fisseni & Fennekels, 1995, 93-20, 153-160).*
- *Ein Beispiel aus der **Pädagogischen Psychologie**: Lernstörungen lassen sich, im Verbund mit anderen Verfahren, durch Verhaltensbeobachtung feststellen:*
 - ⇒ *Informationen können die gestörten Kinder selber liefern, zunächst durch Selbstreport.*
 - ⇒ *Informationen können die Eltern liefern durch Bericht über Verhaltensauffälligkeiten oder neurotische Symptome, ebenso aber auch durch Interaktion mit ihrem Kind unter den Augen des Untersuchers.*
 - ⇒ *Wichtige Informationen liefern schließlich die Beobachtungen, die der Diagnostiker selber bei seinen Untersuchungen sammelt, etwa wenn er die Kinder unter seiner Kontrolle bestimmte Aufgaben bearbeiten lässt. Alle drei Informationsarten können beitragen zur Planung und zur Kontrolle therapeutischer Schritte (Lorenz, 1987).*
- *Ein Beispiel aus der **Verhaltenstherapie**: Welche Störung vorliegt, unter welchen konkreten Bedingungen sie auftritt, wird festgestellt aufgrund exakter Verhaltensbeobachtung und Verhaltensbeschreibung; diese ihrerseits bestimmen wiederum den Plan und den Ablauf einer Therapie wesentlich mit. Ob die therapeutischen Maßnahmen eine Störung gemindert oder gar beseitigt haben, wird schließlich auch entschieden aufgrund sorgfältig dokumentierter Beobachtungen (auch wenn andere Methoden hinzukommen) (Reinecker, 1991).*

- Ein Beispiel aus **Ehe- und Partnerschaftstherapie**: „Neben der umfangreichen Gruppe der Selbstberichtsmethoden und der schier unerschöpflichen Vielfalt der Interviewgestaltung bieten sich zur Bewertung der Gegenwartssituation einer gestörten ehelichen Beziehung Methoden der Verhaltensbeobachtung an. Das ist insbesondere zur Beurteilung der dyadischen Interaktion hinsichtlich ihrer verschiedenen Äußerungsformen erforderlich... Man läßt dazu die Partner über Lösungsmöglichkeiten eines von ihnen vorgetragenen Eheproblems diskutieren, Dieser über etwa 10 Minuten geführte Dialog wird in Abwesenheit des Psychologen per Tonband oder - besser noch - per Video aufgezeichnet. Anschließend wird die Konserve von mindestens einem, nach Möglichkeit unbeteiligten Psychologen anhand einer Verbal Problem Check List geratet... Der Praktiker gewinnt mit dieser Methode nicht nur in relativ kurzer Zeit einen quantitativen Überblick über die Art und Weise, wie die Partner miteinander umgehen. Er erhält auch so etwas wie ein Symptomcluster kommunikativer Besonderheiten in der Ehe, auf die er sich dann in der Intervention konzentrieren kann“ (Scholz, 1987, 87-88).
- Ein Beispiel aus der **Familientherapie**: Um das soziale System „Familie“ zu erfassen, bemühen sich Therapeuten um eine Analyse der Interaktionen zwischen den Mitgliedern der Familie. Dabei gewinnt die Verhaltensbeobachtung an Bedeutung. Grund dafür ist die Suche nach einer ganzheitlichen Erfassung des Familiensystems. Eine Analyse soll sich nicht allein auf Selbstberichte stützen, welche einzelne Familienmitglieder abgeben, sondern soll sich ergeben aus Beobachtungen des aktuellen Verhaltens. Wie die Situationen gestaltet werden können, in der zwischen den Familienmitgliedern Interaktionen provoziert werden, referiert der folgenden Abschnitt (Kötter & Nordmann, 1987, 133).

Provokation von Familieninteraktionen

- „ 1) Die Stimulierungsmethode sollte relevant und interessant für die Partner . . . sein.
- 2) Die Familie sollte . . . reales Problem- und Konfliktlösungsverhalten zeigen können.
- 3) Die Familie sollte allein, ohne einen im Raum anwesenden Beobachter, diskutieren können.
- 4) Die Methode sollte nicht zu großen technischen und zeitlichen Aufwand erfordern...
- 5) Ein Kontakt zwischen Beobachter und Familie sollte schon vor der Beobachtungssituation vorhanden sein; der eigentlichen Diagnostik sollte eine Warming-up-Phase vorangehen.
- 6) Die Untersucher sollten sich über Kontextdifferenzen und deren Effekt auf die Generalisierbarkeit der Ergebnisse klar sein.
- 7) Die Beziehung zwischen Beobachteten und Beobachter und deren Einfluß auf die Beobachtung sollte genügend beachtet werden.“

Quelle: „Beobachtungsmethoden“ von Kötter und Nordmann (1987, 139)

Resümee: Verhaltensbeobachtung kann bei interventiven Maßnahmen wesentliche Aufgaben lösen helfen - bei der Vorbereitung ebenso wie bei der Durchführung und Bilanzierung psychologisch-therapeutischer Eingriffe.

7.6 Vor- und Nachteile der Verhaltensbeobachtung

Verhaltensnähe macht das Instrument der Beobachtung zu einem vorzüglichen Verfahren im Dienst von Diagnostik und Intervention. Diesem Vorzug lassen sich einige Nachteile gegenüberstellen.

Die **Vorteile** seien für die Forschung anders gewichtet als für die Einzelfalluntersuchung:

Für *Forschung und Reihenuntersuchungen* seien zwei Gesichtspunkte genannt:

- ‚Neue‘, wenig beachtete Verhaltensweisen können in die Untersuchungen einbezogen werden.
- Schon ‚bekannte‘ Verhaltensweisen können unter neuen Bedingungen untersucht werden.

In der *Individualdiagnostik* empfiehlt sich die Beobachtung aus verschiedenen Gründen:

- Sie ermöglicht es, eine Mannigfaltigkeit der Verhaltensaspekte zu berücksichtigen.
- Sie bietet die Chance, auch spontanes, nicht vorklassifizierbares Verhalten zu erfassen und für die Diagnose zu nützen.
- Beobachtung hat eine größere Nähe zum individuellen Bios: Ein biographischer Ansatz kann auf Verhaltensbeobachtung nicht verzichten.

Den Vorteilen stehen **Nachteile** gegenüber:

- Die Auswahl der Beobachtungseinheiten unterliegt einer gewissen Willkür: Jeder Untersucher kann jeden Verhaltensaspekt für beobachtenswert erklären.
- Zudem dürfte es schwierig sein, jede Abgrenzung theoretisch zu begründen.
- Die Vergleichbarkeit ist erschwert, wenn Verhaltensbeobachtungen nicht vorklassifiziert (standardisiert) werden.

7.7 Zu den Gütekriterien der Verhaltensbeobachtung

Wenn wir die Verhaltensbeobachtung nach den klassischen Gütekriterien beurteilen, gehen wir von der Annahme aus, daß sich das Modell der klassischen Testtheorie dazu eignet, die Meßqualität von Beobachtungsdaten angemessen abzubilden. Wir setzen Modellverträglichkeit voraus.

*Streng genommen, müßte in jedem Falle geklärt werden, ob die Voraussetzungen der Modellverträglichkeit tatsächlich gegeben sind, beispielsweise ob sich das „beobachtete Verhalten“ als **stabil** charakterisieren läßt (und nur in zufälligen Grenzen fluktuiert).*

Es „wird auch in der Verhaltensbeobachtung die Einhaltung der klassischen Meßgütekriterien, wie Reliabilität und Validität, immer wieder gefordert und empfohlen... Ich erachte eine solche Forderung als äußerst unzweckmäßig und in ihrer Konsequenz als schädlich. Erhebungsinstrumente der Verhaltensbeobachtung sollten ohne die stark belastenden Konzepte der Reliabilität und der Validität auskommen. Denn in ihnen sind sachbezogene statt methodologische Annahmen enthalten. Niemand fordert die Reliabilität und Validität der Längenmessung, wohl aber deren Objektivität. Auch die immer wieder gestellte Frage: *Mißt der Test auch wirklich das, was er messen soll?* ist eine höchst sonderbare Frage. In ihr verbirgt sich ein erkenntnislogisch problematischer Realismus, der danach fragt, ob man mit dem Meterstab auch tatsächlich Länge messen könne. Im Prinzip wird dabei vorausgesetzt, daß wir wissen, wie der betrachtende ‚Realitäts-Ausschnitt wirklich‘ ist“ (Faßnacht, 1995, 219).

Kasten 7-6 gibt einen knappen Überblick, wie Objektivität, Reliabilität und Validität von Beobachtungsdaten beurteilt worden sind.

Kasten 7-6:
Zu den Gütekriterien von Beobachtungsdaten

Objektivität

(Auswerter-Übereinstimmung)

Objektivität von Beobachtungsdaten läßt sich gewährleisten, vor allem im Sinne einer Auswerter-Übereinstimmung (Bungard, 1980, 84; Hasemann, 1983, 472-473).

Zu hoher Auswerter-Übereinstimmung tragen erheblich bei:

- *Schulung der Wahrnehmungsschärfe*: Wiederholte Beobachtung desselben Objektes (Video) mit Diskussion der Fehler;
- *Hervorhebung von Teilaspekten*: den Gesamtgegenstand nach ‚Haupt-‘ und ‚Nebensachen‘ unterscheiden und die Teilaspekte getrennt registrieren;
- *Übung der sprachlichen Darstellung*: eigene Erlebnisse mündlich schildern, kleinere Handlungsabläufe exakt beschreiben.

Beobachterschulung soll „auf die Verbesserung der selbstkritischen Stellungnahme zur eigenen Beobachtungsleistung gerichtet sein“ (Hasemann, 1983, 472).

Reliabilität

Auch die Reliabilität von Beobachtungsdaten läßt sich gewährleisten. Sie wird um so höher ausfallen,

- je präziser die Beobachtungseinheiten *definiert* werden,
- je geringer die *Zahl* dieser Einheiten ist,
- je *konkreter* die Beobachtungseinheit formuliert ist, je weniger sie also zu Abstraktion und Schlußfolgerung nötigt (Bungard, 1980, 82).

Es lassen sich Bedingungen anführen, die eine optimale Zuverlässigkeit ermöglichen, ohne sie zu verbürgen (Hasemann, 1983, 474; vgl. Cranach & Frenz, 1969, 300-304).

Genannt seien:

- Serien *mehrerer* Einzelbeobachtungen,
- im *gleichen zeitlichen* Abstand,
- durch zwei bis vier *geschulte Beobachter*,
- mit *wenigen* Verhaltensdimensionen.

Validität

Als unbefriedigend gilt die Validität von Beobachtungsdaten (Cranach & Frenz, 1969, 305-307; Hasemann, 1983, 476). Die Schwierigkeiten lassen sich zurückführen auf unzulängliche Reliabilität, ebenso auf unbefriedigende Kriterien.

„Typische“ Korrelationen zwischen Prädiktoren und Kriterien liegen, wie Hasemann berichtet (1983, 476), bei $r = 0.30$:

- etwa zwischen einem Verhaltensurteil über einen Probanden (Prädiktor) und der Zustimmung des Beurteilten (Kriterium) oder
- zwischen einem Verhaltensurteil über einen Probanden (Prädiktor) und der Zustimmung von Bekannten / „Experten“ (Kriterium) oder
- zwischen beobachtetem Verhalten eines Probanden (Prädiktor) und seinem Verhalten in einem Test (Kriterium).

„Mit großer Gewissenhaftigkeit angewandt, werden Verhaltensbeobachtung und Rating-Verfahren bis zur Behebung dieses unbefriedigenden Zustandes auch ohne zulängliche Validität als Methoden der Verhaltenserfassung einsetzbar sein“ (Hasemann, 1983, 476).

Als Stütze dieses Satzes der Hinweis: In 75 Prozent aller Untersuchungen, veröffentlicht binnen acht Jahren im „Journal of Applied Behavior Analysis“, wird als Methode der Datenerhebung die Verhaltensbeobachtung verwandt (Hasemann, 1983, 476).

7.8 Zusammenfassung zu Kapitel 7

Verhaltensbeobachtung ist eine Grundlage der Psychologie als empirischer Wissenschaft, also auch ihrer Teildisziplinen Diagnostik und Intervention.

Abgehoben von dieser Basisfunktion, kann Verhaltensbeobachtung als selbstständiges Verfahren neben anderen Verfahren eingesetzt werden, z.B. um situative Bedingungen einer Untersuchung zu registrieren oder andere relevante Erkenntnisse zu vermitteln.

Umschreiben läßt sich Beobachtung als eine besonders aufmerksame Wahrnehmung, die sich auf wohldefinierte Verhaltensausschnitte richtet. Sie schließt Selbst- und Fremdbeobachtung ein. Vorrang hat in der Psychologie die Fremdbeobachtung - so weit möglich, unter Beteiligung *mehrerer* kompetenter Beobachter.

Probleme bereitet die Abgrenzung von Beobachtungseinheiten. Ideal ist der Modus einer Abgrenzung, der sich an einer Theorie orientiert.

Die Abgrenzungsfrage taucht in allen Einteilungen auf. Zwei Einteilungen seien genannt: erstens Beobachtung als systematisches oder unsystematisches Vorgehen; zweitens Beobachtung in natürlichen und in künstlichen Situationen.

Ein Problem eigener Art gibt die Registrierung auf: Verhalten kann mechanisch fixiert (z.B. auf Film oder Tonband) oder aber sprachlich festgehalten werden. In dieser ‚Übersetzung‘ verbirgt sich das brisante Problem, wie weit Sprache

schon das Beobachten beeinflusst und wie weit sie das Beobachtete genau ‚abbildet‘.

In der Diagnostik wird die Verhaltensbeobachtung zweifach angewandt. Als begleitendes Verfahren kann sie den situativen Kontext vieler Verfahren registrieren. Als selbständiges Verfahren kann sie die Ergebnisse vieler Verfahren ergänzen.

Für die Intervention kann die Beobachtung ‚Eingriffsorte‘ identifizieren, kann Interventionsschritte mitdefinieren und das Ergebnis interventiver Maßnahmen registrieren.

Die Vorteile der Verhaltensbeobachtung liegen in der Mannigfaltigkeit der Verhaltensaspekte, die beachtet werden können. - Nachteile können der ‚Willkur‘ einer Abgrenzung von Einheiten entspringen sowie der eingeschränkten Vergleichbarkeit unterschiedlicher Beobachtungen.

In jeder Verhaltensbeobachtung ist mit Fehlern zu rechnen, zusammenfaßbar unter dem Titel der ‚impliziten Persönlichkeitstheorie‘. Den Fehlern ist gemeinsam, daß Sachverhalte nicht festgestellt, sondern aus unbegründeten Vorurteilen heraus nur angenommen werden.

7.9 Kontrollfragen zu Kapitel 7

- Definition.
- Korrelativer Zusammenhang von Selbst- und Fremdbeobachtung.
- Gründe für den Vorrang der Fremdbeobachtung.
- Abgrenzung von Beobachtungseinheiten.
- Klassifikationsmöglichkeiten.
- Anwendung in der Diagnostik.
- Anwendung in der Intervention.
- Fehler.
- Gütekriterien.

8. Kapitel

Gesprächsführung, Exploration, Interview, Anamneseerhebung

„Gespräch“ bezeichnet eine Vorgehensweise der Informationssuche, bei der ein Proband durch gezielte Fragen zu Angaben über sich und sein Umfeld angeregt werden soll. Andere Namen lauten Anamnese, Exploration oder Interview. Die Bezeichnung „Gespräch“ dient als Oberbegriff.

Auf „Gespräche“ ist der diagnostische Prozeß angewiesen. In einem „Gespräch“ schildert der Proband dem Psychologen sein Problem. In einem „Gespräch“ teilt Diagnostiker oder Therapeut dem Probanden seine Befunde und seine Stellungnahme mit.

Denkt der Untersucher - nicht nur unreflektiert, sondern auch bewußt - von einem anthropologischen Ansatz her kann er das Gespräch als zentrales Diagnosticum auffassen. Mit dem Gespräch wählt er jenes Mittel, in dem er den Probanden als Partner anerkennt: als jemanden, mit dem er einen Dialog eingeht. So wird beispielsweise nur im Dialog ein individueller Bios in seiner zeitlichen Erstreckung zugänglich, nur im Dialog läßt sich die subjektive Bedeutsamkeit von Ereignissen erschließen - ein Sachverhalt, der seinerseits bedeutsam werden kann für Diagnostik und Intervention ...

Entscheidend ist dieser Zugang für einen Diagnostiker oder Therapeuten, der Persönlichkeit als Prozeßgestalt interpretiert und darum annimmt, daß ein Individuum sich nur aus seiner Geschichte „erklärt“ oder „erklären läßt“ (Bühler, Ch., 1969, 10; Dailey, 1960, 22; Fisseni, 1987, 249-265; Kelly, 1958, 56; Murray, 1938, 39, 604; Thomae, 1968, 111; Undeutsch, 1983).

Den umfangreichen Stoff gliedern wir in drei Teilkapitel:

- Vorklärunen und Festlegungen (8.1),
- Vorbereitung, Durchführung, Auswertung von Gesprächen (8.2),
- Gütekriterien explorativer Daten (8.3).

Es folgen eine Zusammenfassung (8.4) und eine Reihe von Kontrollfragen (8.5).

8.1 Vorklärungen und Festlegungen

Unter dem Titel der „Vorklärungen und Festlegungen“ behandeln wir drei Themen:

- Abgrenzungen (Definitionsfragen) (8.1.1),
- Klassifikation von Gesprächen (8.1.2),
- Explorative Fragetechniken (8.1.3).

8.1.1 Abgrenzungen (Definitionsfragen)

Gespräch dient hier als ein Oberbegriff. Speziellere Bezeichnungen lauten: Exploration, Interview oder Erhebung einer Anamnese. Zufolge vielfältiger, auch disparater Anwendung hat sich keine einheitliche Terminologie durchgesetzt. Wir suchen nach sprachlichen und inhaltlichen Abgrenzungen.

Anamnese geht zurück auf das griechische Verb ‚anamimnéskein‘, das soviel bedeutet wie ‚erinnern, erwähnen, in Erinnerung rufen‘ (Kaegi, 1904, 57). Das Substantiv ‚Anamnese‘ wurde in die medizinische Fachsprache aufgenommen, es bezeichnet die Krankheitsgeschichte, die der Arzt zu Beginn einer Behandlung erhebt. Von der Medizin übernahm die Psychologie den Begriff, erweiterte aber seinen Bedeutungshof (Kemmler, 1974, 9).

In der Psychologie bezeichnet Anamnese die Erfassung der Biographie eines Menschen. Doch geht es nicht darum, nur die ‚Störungen des Verhaltens‘ zu ermitteln (in Entsprechung zur Krankheitsgeschichte), sondern um eine Beschreibung des gesamten Entwicklungsverlaufes (Kemmler 1974, 9-10).

Exploration leitet sich vom lateinischen Verb ‚explorare‘ her, das soviel bedeutet wie ‚ausforschen, ermitteln, erkunden, untersuchen, prüfen, einer Sache auf den Grund gehen‘ (Blase & Reeb, 1909, 298). Als ‚Gespräch‘ umschreibt Exploration demnach ein Vorgehen, das darauf abzielt, den ‚Subjektiven Lebensraum‘ des Probanden zu ‚erkunden‘.

Interview läßt sich bestimmen als eine „Begegnung von Personen, die sich treffen, um miteinander zu diskutieren, Fragen zu besprechen oder Meinungen auszutauschen“ (a meeting of persons face to face for the purpose of discussion, asking questions and getting opinions: Hornby, Gatenby & Wakefield, 1960, 666).

Einige Autoren, etwa Lehr (1964, 97) und Thomae (1968, 111-112), treffen klare Abgrenzungen:

- Als **Interview** bezeichnen sie ein ‚Gespräch‘, das eher sachbezogen ist, sich also auf neutrale Sachverhalte richtet. Der Interviewer bewahrt die Rolle des Beobachters: Ihn interessiert die Information, nicht die Person des Befragten.

- Von **Exploration** und **Anamnese** sprechen sie, wenn „die Persönlichkeit des Gesprächspartners selbst“ der Gegenstand des ‚Gesprächs‘ ist (Lehr 1964, 98), wenn sich die Informationssuche also deutlicher aufpersönliche (auch intime) Auskünfte richtet. -Exploration meint dabei vorwiegend das Gespräch mit dem Befragten, Anamnese vorwiegend das Gespräch mit Drittpersonen (über die Entwicklungsgeschichte des Probanden).

Die meisten Autoren trennen die Begriffe nicht so eindeutig, vor allem nicht im angelsächsischen Sprachraum. Dort überwiegt die Bezeichnung ‚Interview‘.

- Im Deutschen kommt als vierter Terminus ‚Befragung‘ hinzu.

Wir verwenden die vier Titel (Anamnese, Befragung, Exploration, Interview) **synonym**, bevorzugen aber die beiden Benennungen ‚Exploration‘ und ‚Anamnese‘ in dem von Lehr und Thomae umschriebenen Sinne.

Gemeinsame Charakteristika: Der Prozeß der Informationssuche, den die vier Titel bezeichnen, läßt sich gemeinsam unter vier Perspektiven charakterisieren:

1. Es handelt sich um Informationsvermittlung durch **Wechselrede** zwischen mindestens zwei Personen.
2. Der **Informationsfluß** geht primär **in eine Richtung**: vom Befragten zum Befrager.
3. Zwischen Befrager und Befragtem laufen **Interaktionen** auf unterschiedlichen Ebenen ab (Ulich, D., 1982, 48-50):
 - *Allgemeinpsychologisch* betrachtet, handelt es um den Prozeß eines Informationsaustausches.
 - *Sozialpsychologisch* vollziehen sich unterschiedliche Formen der Kommunikation, von neutralem Kontakt bis hin zu persönlicher Begegnung (erlebt zum Beispiel als Sympathie oder Antipathie).
 - *Lernpsychologisch* finden Prozesse wechselseitiger Verstärkung statt, die als Belohnung oder Bestrafung den Informationsstrom lenken.
 - *Tiefenpsychologisch* ist zu rechnen mit Abwehr- und Übertragungsvorgängen.
4. Das ‚Gespräch‘ wird von seiten des Befragers so angelegt, daß sich die **Auswertung** rational kontrollieren läßt. Gegebenenfalls werden die Informationen in quantifizierte Kategorien übertragen.

Die Gemeinsamkeiten sollten nicht dazu verleiten, die Unterschiede zu übersehen. ‚Gespräche‘ erhalten unterschiedliche Funktionen je nach dem Kontext, in dem sie geführt werden (Schraml, 1964, 871-872). Darauf sei kurz eingegangen!

Zwei Besonderheiten: Mehr beiläufig als ausführlich seien zwei Besonderheiten illustriert, (1) die unterschiedliche Rolle des Befragers in unterschiedlichen Gesprächsformen, (2) drei besondere Spielarten einer Befragung.

Zu (1): Unterschiedliche Rolle des Befragers: Wir erwähnen drei Beispiele: Psychoanalyse, Gesprächs- und Verhaltenstherapie.

In der **Psychoanalyse** läßt sich der Therapeut nur in einem abgeschwächten Sinne als ‚Befrager‘ bezeichnen. Er spielt mehr die Rolle eines schweigenden Zuhörers, eines - der Vergleich sei erlaubt - ‚Katalysators‘. Von sich aus ‚erfragt‘ er keine Sachverhalte, sondern wartet, bis Assoziationen den Probanden dazu bringen, selber Sachverhalte zu ‚offenbaren‘.

In der **Gesprächstherapie** hat der Befrager nur die Aufgabe, die Gefühle des Befragten zu ‚spiegeln‘ und ihm die Erfahrung zu vermitteln, daß er ohne jede Einschränkung aussprechen darf, was er gegenwärtig fühlt und erlebt.

In der **Verhaltenstherapie** ermittelt der Befrager vom Klienten exakt, in welchen Situationen und unter welchen Bedingungen eine Störung auftritt; ebenso präzise konzipiert und bespricht er mit ihm, in welchen Situationen und unter welchen Bedingungen die Störung „korrigiert“ werden kann.

Zu (2): Sonderformen der Befragung: Wir erwähnen drei Beispiele: Tiefen- und Gruppeninterview sowie die schriftliche Befragung.

Das **Tiefeninterview** bestimmt sich durch die Detailliertheit der Befragung. Inhaltlich kann es sich um differenzierte Fragen zum Tagesablauf handeln, aber ebenso um Fragen, die (im Umfeld tiefenpsychologisch orientierter Vorgehensweisen) an unbewußte Prozesse wie Übertragung und Gegenübertragung heranführen sollen.

Gruppeninterviews gestatten es, Probanden zu Interaktionen zu veranlassen, etwa zu einer Gruppendiskussion. Der Diagnostiker kann dann *tatsächliches* Verhalten wahrnehmen und erfassen - Verhalten, das *gezeigt* und nicht nur beschrieben wird.

Die **schriftliche Befragung** wird zuweilen angewandt, wenn eine große Stichprobe interviewt werden soll. Durchführbar ist sie zum einen „als Beantwortung eines persönlich überreichten Fragebogens in Gegenwart einer Aufsichtsperson“ (Anger, 1969, 591), zum anderen auf postalischem Wege (Anger, 1969, 598-591; Wilk, 1975, 187-200): Probanden werden angeschrieben und um Beantwortung von Fragebögen gebeten (vergleiche etwa die Befragung von Giese & Schmidt, 1968, zur Sexualität bei Studenten).

Anwendung: Für die Diagnostik, vor allem die des Einzelfalles, dürften Tiefeninterviews und Gruppengespräche nützlich sein, weniger die schriftliche Befragung. (In der Rolle einer Vorselektion, etwa bei Bewerbern um dieselbe Stelle, können auch schriftliche Befragungen dienliche Informationen liefern.)

8.1.2 Klassifikation von Gesprächen

Das Gespräch läßt sich unterschiedlich klassifizieren. Wir führen nur zwei Unterscheidungen an:

- weiches, neutrales, hartes Gespräch (8.1.2.1),
- standardisiertes, unstandardisiertes,
halbstandardisiertes Gespräch (8.1.2.2).

8.1.2.1 Weiches, neutrales, hartes Gespräch

Eine erste Klassifikation orientiert sich an der Art, wie der Befrager seine Rolle gegenüber dem Befragten festlegt. Nach der *Gestaltung dieser Interaktion* lassen sich drei Formen unterscheiden - weiches, hartes und neutrales Gespräch:

- Im **weichen Interview** strebt der Befrager eine Atmosphäre der Offenheit und Wärme an, die es dem Befragten ermöglicht, sich zu ‚öffnen‘ und individuelles Erleben mitzuteilen.
- Im **neutralen Interview** reagiert der Befrager zurückhaltend auf die Äußerungen des Befragten und bekundet eine Haltung freundlichen Gewährenlassens. Angezielt wird die Kontrolle unerwünschter Einflüsse und eine hohe Vergleichbarkeit der Angaben.
- Im **harten Interview** versucht der Befrager durch Überrumpelung, Einschüchterung oder Provokation die Abwehr des Befragten zu durchbrechen und seine Offenheit zu ‚erzwingen‘.

Hartes und weiches Interview beruhen auf der Annahme, der Befragte sei von sich aus nicht bereit, eine ‚wahre Antwort‘ zu geben, darum müsse sein Widerstand durch eine geschickte Technik überwunden werden.

Anwendung: In Diagnostik und Intervention dürfte der Normalfall das neutrale Gespräch sein. Nur in Sonderfällen sollte sich der Untersucher auf das weiche oder das harte Interview einlassen.

8.1.2.2 Standardisiertes, unstandardisiertes, halbstandardisiertes Gespräch

Eine zweite Klassifikation orientiert sich am Gesprächsinhalt, sie bezieht sich auf den Freiheitsgrad, der dem Befrager und dem Befragten eingeräumt wird bei *Wahl und Gestaltung der Gesprächsthemen*. Unterscheiden lassen sich drei Arten: standardisiertes, unstandardisiertes, halbstandardisiertes Gespräch.

Es handelt sich um die wichtigste Klassifikation von Gesprächen (Anger 1969, 570-574; Guilford, 1959, 154-156; Hron, 1982; Kohli, 1978, 11; Mac-coby & Maccoby, 1965, 39-45; Zetterberg, 1969, 206).

Standardisieren lassen sich zum einen die *Reize* (die Fragen), zum anderen die *Reaktionen* (die Antworten). Am Beispiel des Gespräches veranschaulicht Kasten 8-1 Modelle der Standardisierung.

Kasten 8-1:
Standardisierungsmodelle am Beispiel des Gesprches

Standardisiert knnen sein: - allein der Reiz (die Frage) oder allein die Reaktion (die Antwort), - sowohl der Reiz wie auch die Reaktion .			
		Reaktion (Antwort)	
Reiz (<i>Frage</i>)		<i>Standardisiert</i>	<i>Unstandardisiert</i>
		A <i>Standardisiertes Gesprch</i>	B <i>Halbstandardisiertes Gesprch</i>
	Stan- dardi- siert	Frage: standardisiert Antwortoptionen: standardisiert	Frage: standardisiert Antwort: unstandardisiert
	Unstan- dardi- siert	C <i>Halbstandardisiertes Gesprch</i> Frage unstandardisiert Antwortoptionen: standardisiert	D <i>Unstandardisiertes Gesprch</i> Frage: unstandardisiert Antwort: unstandardisiert

Standardisiertes Gesprch

Im standardisierten Gesprch sind Formulierung und Reihenfolge der Fragen, ebenso die Antwortklassen und Auswertekategorien vollstndig festgelegt. Weder Befrager noch Befragter drfen die Fragen ndern. Der Befrager darf keine anderen als die vorgegebenen Antwortoptionen zulassen, der Befragte darf nur Antworten geben, die sich in die vorgegebenen Antwortoptionen einordnen lassen.

Diese Form des Interviews steht dem Persnlichkeitsfragebogen im engeren Sinne (dem Persnlichkeitstests) sehr nahe.

Welche **Grnde** sprechen fr die Wahl eines standardisierten Interviews?

- 1. Anwendung und Auswertung sind *konomisch* (etwa nach Material und Zeit).
- 2. Die Informationen aus mehreren Interviews lassen sich leicht *vergleichen*.
- 3. Es ist leicht mglich, *Gtekriterien* wie Objektivitt, Reliabilitt und Validitt zu ermitteln.

Es seien auch **Nachteile** genannt: ‚Standardisierung‘ kann dazu fhren, da der subjektive Lebensraum des Individuums unzureichend abgebildet wird. Bei Reihenuntersuchungen und im Rahmen von Forschungsaufgaben darf oder mu der Untersucher diesen Nachteil in Kauf nehmen; der Vorteil der Vergleichbarkeit (der Generalisierbarkeit) von Daten wiegt ihn auf. - In der Einzelfalldiagnostik kann es sein, da diese ‚Allgemeinheit‘ den Zugang zum Kern einer diagnostischen Frage versperrt.

Unstandardisiertes Gespräch

Im unstandardisierten Gespräch bleiben Inhalte und Reihenfolge der Fragen offen, ebenso die Antwortmöglichkeiten und die Art der Auswertung. (Festgelegt sind allenfalls die Themen, die eingeführt werden sollen.) Der individuellen Variation im Informationsaustausch wird Freiheit gewährt, beim Befrager ebenso wie beim Befragten.

Welche **Gründe** sprechen für die Wahl des unstandardisierten Gesprächs?

1. Eine unstandardisierte Exploration läßt sich lebensnäher gestalten, enger orientiert am individuellen Bios.
2. Befrager und Befragter können wichtige *Themen* beliebig weit verfolgen.
3. Für *unterschiedliche Probandengruppen* lassen sich *unterschiedliche Spruchregelungen* aufnehmen. Bedeutungen lassen sich gleich halten, ohne daß die Worte gleich lauten (Bedeutungsäquivalenz statt Wortäquivalenz).

Die unstandardisierte Form schließt auch **Nachteile** ein:

1. Informationen können so ‚individuell‘ gestaltet werden, daß *Vergleiche* (mit anderen Gesprächen) *erschwert* werden oder unmöglich sind.
2. Wichtige *Informationen* können *ausgelassen* (vergessen oder unterschlagen) werden. Der Proband kann seine Schwächen verbergen, seine Stärken ins Licht stellen.

Diese Gefahren machen aufmerksam auf einen entscheidenden Punkt: Das unstandardisierte Gespräch kann nur gelingen, wenn der Befrager gründlich geschult worden ist.

Halbstandardisiertes Gespräch

Zwischen den beiden Polen „standardisiertes“ und „unstandardisiertes“ Interview sind viele Mischformen möglich - die sogenannten halbstandardisierten Gespräche.

Was kennzeichnet das halbstandardisierte Gespräch?

- Wie ein unstandardisiertes Gespräch *gewährt es Freiheit für individuelle Variationen*.
- Wie ein standardisiertes Gespräch *gibt es Strukturen* vor, welche die Interaktion zwischen Befrager und Befragtem kanalisieren, somit auch die Auswertung und viele Vergleiche erleichtern.
- Es erlaubt dem Untersucher, einen Teil der Fragen und der Auswertekategorien *wörtlich vorzuformulieren*.
- Aber ebenso ermöglicht *es ihm, einen Teil der Fragen im Gespräch frei zu formulieren* (und sie später in neu konzipierten Auswertekategorien einzuordnen).

Anwendung: Die drei Gesprächsarten haben zu unterschiedlichen Zielen eine unterschiedliche Affinität:

- In der *Individualdiagnostik* dürften halbstandardisierte Explorationen überwiegen, orientiert an Themen wie:
 - ⇒ biographische Entwicklung,
 - ⇒ Interessen,
 - ⇒ soziale Interaktionen
 - ⇒ usw.
- In der *Gruppendiagnostik* (Forschung oder Reihenuntersuchung) dürfte den standardisierten Gesprächen ein Vorrang zufallen, zufolge der Ökonomie ihrer Anwendung und der Vergleichbarkeit ihrer Informationen.

Nutzung vorliegender Leitfäden

Für alle drei Interviewformen (standardisiert / unstandardisiert / halbstandardisiert) kann der Untersucher aus vorgegebenen Verfahren Fragehilfen entleihen:

- ***Halb- oder vollstandardisierte Leitfäden*** liegen für bestimmte Verhaltensbereiche vor und können übernommen werden, beispielsweise der ‚Diagnostische Elternfragebogen (DEF)‘ von Dehmelt, Kuhnert und Zinn (1981).
- Aus ***Einstellungs- oder Weressenests*** lassen sich Explorationshilfen übernehmen, beispielsweise aus dem ‚Problemfragebogen für Jugendliche‘ von Süllwold und Berg (1967). Einzelne Themenbereiche kann der Untersucher als ‚Leitfaden‘ ausgliedern.
- Es liegen ***Sammelwerke*** vor, welche für unterschiedliche Probleme *Interviewleitfäden* auflisten. Dafür einige Beispiele:
 - ⇒ Mash und Terdal (1980) haben ein „*Kompendium der verhaltenstherapeutischen Diagnostik*“ publiziert, das einen Untersucher bei seiner Interview-Vorbereitung inspirieren kann.
 - ⇒ Eberwein (1992) hat „*Verfahren der Klinischen Psychologie*“ zusammengestellt, deren Lektüre den Entwurf von Leitfäden erleichtern und bereichern kann.
 - ⇒ Hank, Halweg und Klann (1990) haben „*Diagnostische Verfahren für Berater*“ gesammelt, die einen Diagnostiker bei seiner Gesprächsvorbereitung anregen können.
 - ⇒ Scholz (1978, 1987) entfaltet Beispiele, Anregungen, Erfahrungen zur *Erfassung, Besprechung, Behandlung von Partnerschaftsstörungen*.
- ***Gespräche mit Kindern*** erfordern spezielle Formen der Gesprächsführung. Vorschläge dafür finden sich beispielsweise
 - ⇒ bei Deegener (1984) in „*Anamnese und Biographie im Kindes- und Jugendalter*“,
 - ⇒ bei Lohaus (1989) in „*Datenerhebung in der Entwicklungspsychologie*“,
 - ⇒ bei Rauchfleisch (1991) in der Sammlung „*Kinderpsychologische Tests*“,
 - ⇒ bei Yarrow (1960) in „*Interviewing children*“ (vgl. auch Gross, 1985).

8.1.3 Explorative Fragetechniken

Wenn eine Exploration gelingen soll, muß der Befrager die richtigen Fragen stellen. Das Wort ‚Frage‘ hat in diesem Kontext zweierlei Bedeutung. Es bezeichnet erstens Sätze, die grammatikalisch echte Fragen sind („*Haben Sie einen Bruder?* “), zweitens Sätze, die grammatikalisch zwar Feststellungen sind, aber wie Fragen zum Reden anregen sollen („*Ihren Worten entnehme ich, daß Sie einen Bruder haben. Erzählen Sie mir etwas über ihn!*“)

Zur explorativen Fragetechnik finden sich Hinweise bei fast allen Autoren, die über ‚Befragung‘ schreiben, beispielsweise bei: Anger, 1969, 574-588; Cantril & Rugg, 1965, 86-114; Dahmer & Dahmer, 1982, 65-82; Guthke, Böttcher & Sprung, 1991, 94-102; Hron, 1982, 120-123; Lutz, 1978, 37-38; Noelle, 1968, 54-96; Sarges, 1995, 484-486; Scheuch, 1967, 142-143; Tomm, 1994; 180-196; Westhoff & Kluck, 1991, 105-110).

Die diversen Fragetechniken seien unter zwei Perspektiven vorgestellt: Wir *klassifizieren* die Fragen nach ihrer unterschiedlichen Aufgabe (8.1.3.1), und wir zitieren Regeln zur *Formulierung von Fragen* (8.1.3.2).

8.1.3.1 Klassifikation von Fragen

Nach der Rolle, die den Fragen zufällt, unterscheiden wir hier drei Gruppen: erstens funktionale Fragen (8.1.3.1.1), zweitens formale Fragen (8.1.3.1.2) und drittens Suggestivfragen (8.1.3.1.3).

8.1.3.1.1 Funktionale Fragen

„Funktionale Fragen“ sollen größere Einheiten des Gespräches steuern, sie haben Gelenk- oder Schaltfunktion. Genannt seien drei Arten:

- Kontakt- oder Einleitungsfragen,
- Überleitungs- oder Übergangsfragen und
- Kontrollfragen.

Kontakt- oder Einleitungsfragen haben Eisbrecherfunktion. Sie sollen Vertrauen herstellen und es dem Befragten ermöglichen, in die diagnostische Situation einzutreten. Als Regel gilt, „nicht mit der Tür ins Haus zu fallen“, sondern am Beginn eines Gespräches tabufreie Themen anzusprechen (die Herkunft, die Wetterlage, die Tages- oder Jahreszeit) und von solchen Randfragen zur Kernfrage überzuleiten.

Überleitungs- oder Übergangsfragen sollen von einem Thema zum anderen führen. Eine Kunst ist dies vor allem, wenn ein redefreudiger Proband auf ein Thema seiner „Herzensergießung“ getroffen ist und sich in seinem Redefluß nicht unterbrechen lassen will.

Möglichkeiten der Überleitung bieten sich, indem der Befrager einen bisher besprochenen Abschnitt zusammenfaßt und ein neues Thema anschlägt, etwa in dem Sinne: „Wir haben bisher über Ihre Berufswahl gesprochen. Wir wenden uns einem verwandten Gebiet zu: Welche Fähigkeiten bringen Sie mit, um den gewählten Beruf auszuüben?“

Kontrollfragen sollen dazu beitragen, Unklarheiten aufzuhellen, etwa scheinbare Widersprüche aufzuklären oder wirkliche Widersprüche aufzudecken. Kontrollfragen setzen, bei Formulierung und ‚Äußerung‘, viel Fingerspitzengefühl voraus; sonst können sie auf den Probanden wirken wie ‚Machtdemonstrationen‘ des Befragers.

8.1.3.1.2 Formale Fragen

„Formale Fragen“ beziehen sich auf einzelne Bereiche des Gesprächs. Sie legen fest, in welcher ‚Form‘ der Befragte diese Bereiche darstellen oder schildern sollte (Cannell & Kahn, 1968, 552-571; Maccoby & Maccoby, 1965, 46-56). Wichtig sind vor allem drei Gruppen:

- offene und geschlossene
- direkte und indirekte Fragen
- zirkuläre Fragen

Offene und geschlossene Fragen

Die Einteilung in „offene“ und „geschlossene“ Fragen betrifft die Antwortoptionen (Westhoff & Kluck, 1991, 108).

Bei *geschlossenen* Fragen gibt der Untersucher (sowohl die *Frage* als auch) die *Antwortoptionen* vor, im Extremfall braucht der Befragte nur mit Ja oder mit Nein zu antworten. - Bei *offenen* Fragen gibt der Befrager zwar die Fragen vor, überläßt die Antwortform jedoch dem Probanden.

Geschlossene Frage - Beispiele:

Frage: *Wie verbringen Sie Ihre Freizeit?*

Ich nenne drei Tätigkeiten.

Antwortoptionen:	<i>Sport</i>	<i>Ja ()</i>	<i>Nein ()</i>
	<i>Lektüre</i>	<i>Ja ()</i>	<i>Nein ()</i>
	<i>Tätigkeit in einem Verein</i>	<i>Ja ()</i>	<i>Nein ()</i>

Offene Frage - Beispiele:

Wie verbringen Sie Ihre Freizeit?

Was haben Sie am letzten Wochenende getan?

Der Vorteil der offenen Fragen entspricht dem der unstandardisierten Exploration: Sie bewegen den Probanden eher dazu, ausführlich über sich und sein Erleben zu reden, sie lassen ihm den größeren Freiraum. - In der Individualdiagnostik dürften daher offene Fragen eine wichtigere Rolle spielen als ge-

schlossene (vor allem zur Ermittlung des eigentlichen Anlasses der Untersuchung).

Direkte und indirekte Fragen

„Besondere Probleme ergeben sich bei dem Versuch, Sachverhalte zu erfragen, über die der Proband entweder nicht offen sprechen **will** oder über die er beim besten Willen keine gültigen Auskünfte geben **kann**“ (Anger, 1969, 581). In solchen Fällen kann eine andere Art von Fragen weiterführen: sogenannte indirekte Fragen.

Die Einteilung in „direkte“ und „indirekte“ Fragen bezieht sich auf den **Gegenstand** des Gespräches (Westhoff & Kluck, 1991, 109).

Direkte Fragen benennen den Gegenstand selber, auf den sie zielen. *Indirekte* Fragen benennen das Umfeld, dem der Fragegegenstand zugehört; aus dem Umfeld sollen sie das Gespräch zum angezielten Gegenstand hinführen.

Direkte Frage - Beispiele:

Darf ich Sie bitten, mir etwas über Ihre Freunde und Bekannten zu erzählen?

Ich würde gerne von Ihnen hören, wie weit Sie mit Ihrer Arbeit zufrieden sind!

Indirekte Frage - Beispiele:

Erzählen Sie bitte, was Sie am letzten Wochenende getan haben.

Erzählen Sie mir bitte, wo Sie Ihren letzten Urlaub verbracht haben.

Beschreiben Sie mir bitte Ihren Arbeitsplatz.

In beiden Beispielen will der Befrager etwas über die Kontakte und die Berufszufriedenheit des Befragten erfahren: bei der direkten Frage, indem er sich geradewegs nach Freunden und Berufszufriedenheit erkundigt; bei der indirekten Frage, indem er einen Umweg einschlägt (zu Wochenende, Urlaub, Arbeitsplatz) und hofft, daß der Befragte das angezielte Thema selber berührt.

Hinweis: Indirekte Fragen lassen vielerlei Abwandlungen zu, etwa

- projektive Fragen,
- Vervollständigung von Argumenten,
- Wortassoziationen
- usw.

Beziehung zu offenen und geschlossenen Fragen:

- *Direkte* Fragen sind mit *geschlossenen* Fragen verwandt, welche die Antwort vorgeben. Direkte Fragen geben nicht die Antwort vor, sondern nur den *Gegenstand*, auf den sich die Antwort beziehen soll.
- *Indirekte* Fragen sind mit *offenen* Fragen verwandt, welche die Antwort freistellen. Indirekte Fragen benennen den Bereich, über den der Proband sprechen soll und von dem er dann (so hofft der Befrager) zum eigentlich angezielten Gegenstand übergehen wird (Maccoby & Maccoby, 1965, 52-55).

Zirkuläre Fragen

Zirkuläre Fragen sollen erschließen, welche Vorstellungen verschiedene Personen übereinander entwickelt haben. Die beteiligten Personen werden ‚reihum‘ - gleichsam in einem Zirkel - befragt, wie sie über andere Personen denken und fühlen. „Mit diesen Fragen sollen zirkuläre Muster enthüllt werden“ (Tomm, 1994, 182).

Demgemäß hat diese Fragengruppe ihren Ort in Explorationen, in denen soziale Systeme erforscht werden, etwa in einem klinischen Interview der Familientherapie oder - zur Aufdeckung von Konfliktmustern - im Gespräch mit Mitarbeitern einer Verwaltungseinheit.

Beispiel: *„In einer großen öffentlichen Verwaltung ... bestand das typische und problematische Konfliktregelungsmuster darin, neue formale Regelungen auszuarbeiten. So wurden zwischenmenschliche Spannungen, aktuelle Rivalitäten und traditionsreiche Konkurrenzhaltungen nicht angesprochen, sondern um sie herum ein offizielles Netz von Regelungen, Dienstanweisungen und Aktenvermerken angelegt, bis der Gesamtapparat nahezu handlungsunfähig, zumindest aber hochgradig ineffektiv war: Ein Großteil der Verwaltungsenergie war dadurch gebunden, mit Macht den Deckel auf einem Topf zu halten, über dessen brodelnden Inhalt in der Kantine und auf den Fluren - also im informellen System - ganz offen gesprochen wurde“ (Becker & Langosch, 1990, 89).*

Ein Interview, das zirkuläre Fragen einsetzt, kann dazu beitragen, ein solches konflikthafte Beziehungsnetz aufzudecken.

Anwendung: *In der Diagnostik sind alle drei Frageklassen nützlich.* - Für Forschung und Reihenuntersuchung gilt: Geschlossene und/oder direkte Fragen führen das Gespräch rascher weiter, vereinheitlichen auch die Informationen. Sie bringen aber die Gefahr mit sich, ‚an der Oberfläche zu bleiben‘. - Für die *Individualdiagnostik* gilt: Offene und/oder indirekte Fragen können Verhaltensbereiche erschließen, über die der Befragte nur zögernd Auskunft geben will oder geben kann, erschweren aber Generalisierungen. - Bei der Diagnostik sozialer Systeme dürften zirkuläre Fragen dazu beitragen, die Verflechtung zwischenmenschlicher Beziehungen aufzudecken, in ihnen Problemursachen zu erkennen und aus ihnen Vorschläge für Verhaltenskorrekturen abzuleiten.

8.1.3.1.3 Ein Sonderproblem: Suggestivfragen

Die Bezeichnung Suggestivfragen bezieht sich auf den Einfluß, den der Befrager auf den Befragten ausübt - mit dem Ziel, eine bestimmte Antwort zu provozieren. Die Suggestivfrage legt dem Probanden eine bestimmte Antwort nahe, sie verletzt die Neutralität (Anger, 1969, 577; Dahmer & Dahmer, 1982, 73; Noelle, 1968, 52; Scheuch, 1967, 143).

Beispiel: „Meinen Sie nicht auch, daß ...“ - „Nicht wahr Sie waren gestern im Kino?“ - „Sie haben gesagt: Ihre Freundin war ohne Sie im Kino. Merken Sie nicht auch, daß Sie eifersüchtig sind?“

Suggestionen wird der Untersucher kaum vermeiden können. Explizite Suggestionen sollten die Ausnahme bleiben. Zur Erschütterung der Abwehr kann Suggestion zuweilen nützlich sein. Aber sie ist wie ein Gift, das, in kleinen Portionen gereicht, heilen kann, das aber, in großer Dosis gegeben, tödlich wirkt. Pointiert gesagt: ein Gespräch, das auf Suggestion beruht, ist diagnostisch wertlos.

8.1.3.2 Formulierung von Fragen

Die Exploration sollte nicht als Frage-Antwort-Spiel ablaufen. Zu schnell fühlt der Befragte sich bedroht und richtet seine Antworten danach aus, wie er die Sanktionsmacht des Befragers einschätzt. Diese Gefahr ist besonders groß, wenn für den Probanden etwas auf dem Spiele steht, etwa bei Bewerbergesprächen oder Eignungsuntersuchungen. Fragen sollte man darum ergänzen durch Feststellungen, die das Gespräch auflockern.

Wir erwähnen einige Vorschläge, sie zielen auf ein einziges Anliegen: dem Probanden die Antwort zu erleichtern, der Befragung ihren bedrohlichen Charakter zu nehmen, ohne den Antwortweg (im Sinne einer Suggestion) vorzubahnen.

Solche Regeln sind nicht blind zu befolgen, sie sind immer auszulegen mit Blick auf den Gesprächspartner: In einer Exploration mit einem Fernfahrer; der seinen Führerschein ‚wegen Alkohols am Steuer‘ verloren hat, muß ‚Einfachheit der Formulierung‘ etwas anderes bedeuten als bei der Beratung eines Apothekers, der ‚Erziehungsschwierigkeiten‘ hat mit seiner 15jährigen Tochter deren Punk-Manieren er nicht erträgt.

Einfache Formulierungen suchen

Die Regel einfacher Formulierung empfiehlt, eine Sprachebene zu suchen, auf der sich beide Gesprächspartner ‚zu Hause fühlen‘. Der Regel widerspricht der Gebrauch einer Fachsprache, die einer der Partner nicht beherrscht, ihr widersprechen komplizierte Satzkonstruktionen:

- **Kompliziert:**

Können Sie mir bitte die Motivation beschreiben, warum Sie das Psychologiestudium wählen wollen, war es eine altruistische oder eine egoistische Motivation?

- **Einfacher:**

Warum wollen Sie Psychologe werden?

Bitte erklären Sie mir warum Sie Psychologe werden wollen!

Kurze Sätze bilden

Kürze der Fragen erleichtert ein rasches Verständnis. Teilstücke der Frage werden nicht so leicht überhört:

- **Lang:**

Können Sie mir sagen, wann Ihr Berufswunsch, Psychologie zu studieren, entstanden ist, beziehungsweise wann Sie sich seiner zum ersten Mal bewußt geworden sind?

- **Kurz:**

Wann haben Sie zum ersten Mal daran gedacht, Psychologie zu studieren? In welchem Alter kam Ihnen die Idee, Psychologe zu werden? Wie alt waren Sie, als Sie zum ersten Mal daran dachten...?

Eindeutige Formulierungen suchen: Darum keine Doppelfrage, keine Doppelverneinung

Wieder geht es darum, den Informationsaustausch zu erleichtern, dem Partner das Raten zu ersparen. Zwei Beispiel, die den Informationsfluß erschweren, sind Doppelfrage und Doppelte Verneinung.

Eine *Doppelfrage* bringt mindestens zwei Objekte A und B ins Spiel. Damit verursacht sie Unklarheit über den Gegenstand der Frage: Wird nach A oder wird nach B gefragt?

- **Doppelfrage:**

Wissen Sie, ob Sie eher Psychologie studieren wollen, um anderen zu helfen, oder haben Sie eher an das hohe Honorar gedacht, das Sie später als Therapeut beziehen?

- **Einfacher:**

Wie weit hat der Gedanke Ihre Berufswahl mitbestimmt, anderen Menschen zu helfen?

Welche Bedeutung hat es für Sie, daß Sie einen Beruf gewählt haben, in dem man viel Geld verdienen kann?

Eine *Doppelte Verneinung* fordert dem Hörer eine Transformation ab. Zwei Negationen muß er in eine Affirmation umwandeln. Diese Aufgabe belegt Speicherplatz im Gedächtnis und verlängert die Reaktionszeit.

- **Doppelte Verneinung:**

Würden Sie nicht auch von sich sagen, daß Sie kein purer Altruist sind? Könnten Sie nicht ausschließen, daß keine andere Motive Sie beeinflußt haben?

- **Positive Version:**

Anderen Menschen zu helfen: Wie weit hat dieser Gedanke Ihre Berufswahl mitbestimmt?

Welche anderen Motive haben Ihre Berufswahl beeinflußt?

Komplexe Sachverhalte in Einzelfragen zerlegen

Den Partner überfordern Fragen, die seinem Gedächtnis zuviel zumuten. Komplexe Sachverhalte sollte man in Einzelfragen zerlegen:

- **Überforderung:**

Schildern Sie mir bitte die einzelnen Bewältigungsstrategien je nach psychologischer Teildisziplin beziehungsweise je nach sozialer Situation: also, Sie haben unterschiedliche Fächer zu belegen in Ihrem Studium: Statistik, Allgemeine Psychologie, Entwicklungspsychologie, Persönlichkeitstheorie, Sozialpsychologie: bitte geben Sie an, welches Fach schwer und welches leicht für Sie ist und was Sie tun, wenn es Ihnen schwer fällt je nach Fach.

- **Einfacher:**

Sie haben sich in vier Fächern auf eine Prüfung vorbereitet: in Allgemeiner Psychologie, in Entwicklungspsychologie, in Persönlichkeitstheorie und in Sozialpsychologie: welches Fach machte Ihnen Schwierigkeiten?

- Pause: Antwort des Probanden abwarten! -

Was tun Sie bei Schwierigkeiten?

Gegebenenfalls kann der Befrager umfangreiche Themen in Teilbereiche aufgliedern.

Beispiel: Wenn der Untersucher etwa erfahren möchte, wie weit das ‚soziale Engagement‘ des Befragten geht, könnte er das Thema einführen mit den Worten:

„Ich nenne Ihnen einige Tätigkeiten, die ich als Beispiel verstehe:

- *In einem Krankenhaus helfen.*

- *Essen auf Rädern ausfahren.*

- *Ausländerkindern bei Hausaufgaben helfen.*

- *Jugendliche in ein Ferienlager begleiten.*

- *Dienste übernehmen beim Roten Kreuz.*

Jetzt meine Bitte: Nennen Sie mir Tätigkeiten dieser Art, die Sie übernommen haben!“

An die Erfahrung des Probanden anknüpfen

Fragen, die an Erfahrungen des Befragten anknüpfen, erleichtern es ihm, Informationen aufzunehmen und einzuordnen. Anknüpfen können Fragen an Orte, Zeiten oder Zusammenhänge, die der Proband kennt.

Beispiele:

Der Untersucher will etwas wissen über das Freizeitverhalten.

- **Frage → Uneingebettet:**

Wie verbringen Sie Ihre Freizeit?

- **Frage → Eingebettet:**

Wo haben Sie Ihren letzten Urlaub verbracht?

Wann sind Sie morgens aufgestanden?

Was gab es zum Frühstück?

Der Untersucher will etwas wissen über das Verhältnis zum Vorgesetzten.

- **Frage → uneingebettet:**

Wie ist Ihr Verhältnis zu Ihrem Chef?

- **Frage → eingebettet:**

Wann sind Sie zum letzten Mal verspätet zum Dienst gekommen?

Wie hat Ihr Chef reagiert?

Sich mit eigenen Reflexionen zurückhalten

Wenn der Befrager seine eigenen Reflexionen einführt, könnte er Gefahr laufen, die Gesprächsdynamik von der inhaltlichen Ebene auf die Beziehungsebene zu verschieben - bis hin zur Gefahr einer persönlichen Auseinandersetzung. Darum der Ratschlag, sich mit eigenen Reflexionen zurückzuhalten!

Beispiele:

- **Frage in Reflexionen eingebettet:**

Was meinen Sie zu meinem Vorschlag, Ihre Motivationsfragen zu klären, indem wir gemeinsam die Entwicklung Ihres Berufswunsches verfolgen?

- **Neutrale Formulierung:**

Würden Sie bitte den Weg beschreiben, auf dem Sie Ihren Beruf gefunden haben.

Auch dann, wenn Dinge zur Sprache kommen, welche die persönlichen Werte des Befragers berühren, sollte er neutrale Formulierungen wählen, etwa wie: „Es gibt Leute, die sagen...“

Resümee: Worum geht es bei allen Hinweisen zur Formulierung von Fragen? Allein um das Anliegen, dem Befragten den Weg zur ‚wahren Information‘ zu erleichtern.

8.2 Zur Praxis: Vorbereitung, Durchführung und Auswertung von Gesprächen

Teilkapitel 8.2 behandelt folgende Themen:

- Vorbereitung von Gesprächen (8.2.1),
- Durchführung von Gesprächen (8.2.2),
- Auswertung von Gesprächen (8.2.3),
- Beitrag zu Diagnostik und Intervention (8.2.4),
- Fehlertendenzen (8.2.5).

8.2.1 Vorbereitung von Gesprächen

Ein Gespräch mit dem Probanden sollte vorbereitet sein. Stichworte einer Vorbereitung könnten lauten:

- Leitidee und Hauptthemen aufschlüsseln,
- Gesprächsart und Frageklassen wählen,
- Rahmenbedingungen regeln,
- Kontrollen vorsehen.

Leitidee und Hauptthemen aufschlüsseln

Ein Gespräch hat ein Ziel - Umschreibbar in einer Leitidee (*etwa Hilfe bei der Berufsfindung*). Dieser Hauptgedanke zerlegt sich in Einzelfragen (Fragen *etwa nach Berufsmotivation oder nach berufsbezogenen Fähigkeiten und Fertigkeiten*). Münden muß die Aufgliederung bei konkreten Fragen, die in einen Leitfaden passen.

Anschauliche Hinweise zur Erstellung und Verwendung eines Leitfadens finden sich (um nur wenige Namen zu nennen) bei Hron (1984, 124), bei Westhoff und Kluck (1991, 98-105), bei Wittkowski (1994, 42-50).

Leitidee und relevante Themen festlegen: Vor dem Gespräch mit einem Probanden sollte der Befrager die Hauptthemen festlegen, auf die er eingehen will - am besten schriftlich:

- **Vor** einem (ersten) Kontaktgespräch kann diese Abgrenzung nur auf Vermutungen hin geschehen, zum Beispiel aufgrund telefonischer Auskünfte oder früherer Erfahrungen in ähnlichen Fällen.
- **Nach** einem Kontaktgespräch lassen sich die Themenbereiche genauer festlegen; es lassen sich dem Gespräch auch gezieltere Hypothesen zugrundelegen.

Wählen wir als Beispiel die Leitidee einer „*Hilfe bei der Berufssfindung: Psychologiestudium*“. Unter dieser Leitidee stellt der Untersucher Themen zusammen, über die er mit dem Probanden sprechen will.

Welche Themen kommen in Betracht?

- Höhe der Abiturnote,
- Alter, in dem eine Neigung für das Psychologiestudium erwachte,
- Vorstellungen von den Aufgaben eines Psychologen,
- Aussagen zum eigenen Arbeitsstil,
- Vertrautheit mit eigenen Stärken und Schwächen,
- Antizipation und Wertung von Prüfungssituationen,
- Interesse für psychologie-relevante Fächer,
- Interesse für benachbarte naturwissenschaftliche Fächer,
- Einstellung zum Psychologiestudium,

- Einstellung zum Beruf des Psychologen
- usw.

Da nicht anzunehmen ist, daß alle relevanten Themen ‚entdeckt‘ wurden, ist die Aufzählung als offene Liste zu betrachten, nicht als abgeschlossener Katalog.

Themen in eine schlüssige Abfolge bringen: In der Regel ist eine erste Themensammlung nicht systematisch gegliedert. Der Untersucher sollte sie so anordnen, daß er leicht von einem Thema zum anderen weitergehen kann.

Bleiben wir bei dem Beispiel „Berufswahl Psychologiestudium“ und ordnen wir die Liste nach Thematiken, die einander ‚benachbart‘ sein dürften:

- *Am Anfang könnten Fragen stehen, wie der Proband auf die Idee seiner Berufs- und Studienwahl gekommen sei.*
- *Benachbart lägen dann Fragen der Einstellung zu Studium und späteren beruflichen Aufgaben.*
- *Von dort ließe sich weitergehen zu einer Einschätzung eigener relevanter Fähigkeiten und Interessen.*
- *Vom Leistungsbereich, der Frage nach den ‚Fähigkeiten‘, könnte das Gespräch sich der Bedeutung sozialer Kontakte und ihrer Abstimmung auf die Arbeitsanforderungen zuwenden.*
- usw.

Genauso berechtigt ist jede andere Reihenfolge, welche die Themen ‚sinnvoll‘ miteinander vernetzt.

Hauptthemen aufschlüsseln in Teilthemen: Wie lassen sich die Einzelthemen umsetzen in Fragen, die dem Probanden gestellt werden? Ein erster Schritt kann darin bestehen, die *Hauptthemen in Teilthemen aufzuschlüsseln*. - Für ein Einzelthema ‚Arbeitsstil und Arbeitsverhalten‘ bringt Kasten 8-2 eine solche Aufschlüsselung.

Kasten 8-2:
Aufschlüsselung des Themas ‚Arbeitsstil und Arbeitsverhalten‘

<p><i>Weniger wichtig erscheint es uns, zu erfahren, wie im einzelnen das Arbeitsverhalten eines Probanden beschaffen ist. Wichtiger ist es, zu erkunden, ob der Proband über sein Arbeitsverhalten Rechenschaft geben kann.</i></p>			
<p>„Arbeitsstil und Arbeitsverhalten“ Wir gliedern vier Bereiche aus.</p>			
Arbeitsstil	Lerntechniken	Kräfteeinteilung	Durchhaltekraft
Kann der Proband sein Arbeitsverhalten beschreiben?	Kennt er Lerntechniken? Wann setzt er sie ein?	Kann er seine Kräfte abstimmen auf Anforderungen?	Hat er Arbeitsvorhaben durchgesetzt gegen „Widerstand in der eigenen Person“ oder „ungünstige äußere Umstände“?
Hält er seinen Arbeitsstil für elaboriert, für erfolgreich?	Kann er von Mitschülern Techniken übernehmen?	Wie teilt er Arbeit, Pausen, Muße, Schlaf ein?	

Die Aufschlüsselung in Teilthemen führt noch nicht zu Fragen, die dem Probanden gestellt werden könnten.

Nehmen wir die erste Frage in unserem Beispiel: „Kann der Proband sein *Arbeitsverhalten beschreiben*?“ - Selbst wenn die Frage aus der dritten in die zweite Person umgewandelt wird und nun lautet: Können *Sie Ihr Arbeitsverhalten beschreiben*? könnte der Proband sich mit einem einfachen Ja oder Nein begnügen. Der Untersucher will aber mehr erfahren als ein Ja oder Nein.

Eine Aufteilung in Teilthemen ist nur ein erster Schritt auf das Ziel hin, ‚angemessene‘ Fragen zu formulieren.

Gesprächsart und Frageklassen wählen

Hat der Untersucher eine schlüssige Themenfolge gefunden, muß er sich für eine Gesprächsart entscheiden. Diese Entscheidung legt mit fest, welche Frageklassen vorzusehen sind und welche Möglichkeiten einer Aufzeichnung und einer Auswertung in Betracht kommen.

Wir bleiben bei dem Hauptthema der „Berufsfindung“ und Schlüsseln *ein* Teilthema weiter auf: „den Arbeitsstil und das Arbeitsverhalten“.

Explorationsart festlegen und Fragen für einen Leitfaden formulieren: Für das Gespräch wählen wir eine halbstandardisierte Form. Damit ist zugleich festgelegt: Fragen müssen zum Teil ‚geschlossen‘ formuliert werden, zum Teil ‚offen‘ bleiben.

Für das Teilthema „Arbeitsstil und Arbeitsverhalten“ seien einige Fragen entworfen, die in einen Leitfaden eingehen könnten. Kasten 8-3 bringt das Ergebnis.

Kasten 8-3:

Fragen für einen Leitfaden zu dem Thema ‚Arbeitsstil und Arbeitsverhalten‘

*Frage 1 und 3 sind gedacht als offen,
Frage 2 und 4 als geschlossen.*

Fragen

1. Mich interessiert, wie Sie mit Ihren Arbeiten zurechtgekommen sind. Denken Sie bitte an Ihre Jahre in der Oberstufe: Beschreiben Sie mir, wie haben Sie Ihre Arbeit für die Schule organisiert? (*Den Probanden reden lassen!*)
2. Sagen Sie mir bitte auch: Für wie effektiv halten Sie Ihre Art, die Arbeit zu organisieren?
*EINSTUFUNG: Der Proband beschreibt seinen Arbeitsstil als:
sehr ineffektiv 1 - 2 - 3 - 4 - 5 sehr effektiv.*

3. Ich möchte Sie nach einer Einzelheit fragen. Sie wissen, was man unter Lerntechniken versteht. Ich meine damit Hilfen wie: Textstellen unterstreichen, ‚Eselsbrücken‘ bauen, Karteikarten anlegen. *-Jetzt meine Bitte:* Erzählen Sie mir, welche Lerntechniken haben Sie selber verwandt? *(Den Probanden erzählen lassen!)*
4. Wir greifen eine einzelne Lerntechnik heraus, nämlich... Beschreiben Sie mir diese Technik mehr im einzelnen!
EINSTUFUNG: Der Proband schildert eigene Lerntechniken:
sehr undeutlich 1 - 2 - 3 - 4 - 5 sehr deutlich.
5. u.s.w.

Für eine angemessene Aufzeichnung sorgen: Wie soll eine Exploration festgehalten werden? Diese Frage schließt die Frage ein, wie ein Gespräch ausgewertet werden soll - es ist der nächste Punkt, der zu bedenken ist.

Drei übliche Formen der Registrierung seien erwähnt:

- **Mechanische Registrierung** meint eine Fixierung auf Tonband oder Film/Video.
- **Verbale Registrierung** bezieht sich auf Mitschrift oder Nachschrift:
 - a) *Mitschrift:* Der Gesprächsinhalt wird während der Exploration in einer freien Beschreibung mitnotiert. (Gefahr: Der Befrager ‚bemerkt‘ und notiert die Inhalte selektiv.)
 - b) *Nachschrift:* Der Gesprächsinhalt wird nach der Exploration in einer freien Beschreibung rekapituliert. (Gefahr: Der Befrager prägt sich den Inhalt selektiv ein und reproduziert die behaltenen Inhalte noch einmal selektiv.)
- **Numerische Registrierung** bezeichnet Skalen, auf denen die Gesprächsinhalte klassifiziert werden.

Alle drei Arten der Registrierung (mechanische, verbale, numerische) lassen sich einzeln anwenden, alle drei lassen sich aber auch kombinieren.

Die Auswertung vorplanen: Schon bei der Vorbereitung sollte der Untersucher festlegen, wie er ein Gespräch auswerten will. *Das Ziel der späteren Verwendung legt die Art der Auswertung fest.* Die Vorüberlegung könnte zwei Modellen gelten:

- *Soll das Gespräch im Originalfestgehalten werden?* Wenn ja, ist die Frage der Registrierung mitbetroffen: Der Gesprächsverlauf muß auf Film oder Tonband mitgeschnitten werden. (Diese Registrierung kann erhebliche Bedeutung gewinnen, etwa im forensischen Bereich bei Zeugenaussagen.)
- *Soll das Gespräch gekürzt und zusammengefaßt werden?* Wenn ja, ist wiederum die Frage der Registrierung mitbetroffen: Der Gesprächsverlauf muß so genau festgehalten werden, daß eine inhaltsgetreue Zusammenfassung möglich ist. - Mitschnitt auf Film oder Tonband bleibt das Ideal. Mit- oder Nachschrift bleiben Notlösungen.

Abschnitt 8.2.3 bespricht die Auswertung im einzelnen (S.236).

Rahmenbedingungen regeln

Mit dem Titel „Rahmenbedingungen“ spielen wir an auf drei Größen an: räumlich-soziales Umfeld, Zahl und Zeitpunkt von Gesprächen.

Das räumlich-soziale Umfeld vorstrukturieren: Unmittelbar beeinflusst werden Befrager und Befragter von der Atmosphäre des Raumes, in dem *sie* miteinander sprechen, von der Sitzordnung, von der Zahl der beteiligten Personen und von der Dauer eines Gesprächs.

- Was den **Raum** angeht, so wird der Untersucher oft keine Wahl haben. Er sollte aber wissen, daß zum Beispiel Größe, Ausstattung, Zimmertemperatur den Ablauf der Exploration beeinflussen können.
- Was die **Sitzordnung** betrifft, so sollten (wenn vermeidbar) Befrager und Befragter einander nicht am Schreibtisch gegenüber sitzen, sondern, wenn vorhanden, an einem Tisch Platz nehmen.
- Was die **Zahl der Beteiligten** angeht, so sind Einzel- oder Gruppengespräche möglich, je nach der Absicht, die im Gespräch verfolgt wird:
 - ⇒ Einzelgespräche sind unerlässlich, wenn es um persönliche, gar um intime Informationen geht.
 - ⇒ Gruppengespräche erscheinen angemessen, wenn es um Ausgangsfragen oder um Schlußberatungen geht, sofern der Proband nicht allein und persönlich betroffen ist. An Gruppengesprächen können beteiligt sein mehrere Befrager und mehrere Probanden in unterschiedlicher Besetzung (mehrere Befrager und ein Proband oder mehrere Probanden und ein Befrager usw.).
- Ein Gespräch sollte **nicht länger als 60 bis 90 Minuten** dauern.

Mehrfache Gespräche einplanen: Wenn möglich, sollte der Befrager mehrfache Explorationen ansetzen. Dabei können im zweiten oder im dritten Gespräch Informationen als Heurismen dienen, die aus anderen Verfahren stammen, beispielsweise aus Fragebogen oder Verhaltensbeobachtung, aus früheren Gesprächen oder aus projektiven Verfahren.

Das Vorgehen, mehrere Gespräche anzusetzen, empfiehlt sich vor allem in der Einzelfalldiagnostik. - In der Forschung oder bei Reihenuntersuchungen schränkt sich die Zahl möglicher Gespräche, schon aus ökonomischen Gründen, ‚von selber‘ ein.

Den günstigen Zeitpunkt auswählen: Je nach der Absicht, die mit einem Gespräch verbunden wird, läßt sich der Zeitpunkt ansetzen, zu dem das Gespräch stattfindet:

- Gespräche, die das diagnostische Problem erst entwickeln sollen, gehören an den Beginn der Untersuchung. Zweite oder dritte Gespräche sollten so plaziert werden, daß sie es dem Befragten erleichtern, sich mitzuteilen.
- In manchen Fällen kann es geraten sein, ein Gespräch ans Ende der Untersuchung zu verlegen, zum Beispiel auf den Abend, um nach einer Phase

der „Erwärmung“ oder auch der „Ermüdung“ Abwehrhaltungen besser ‚aufweichen‘ oder ‚aufbrechen‘ zu können.

Kontrollen vorsehen

Das Gespräch als soziale Interaktion ist anfällig für Verzerrungstendenzen. Darum der Rat an den Untersucher, seine Tätigkeit immer wieder selber zu kontrollieren und sie der Kontrolle von Kollegen zu unterwerfen!

Mit eigenen Fehlern rechnen: Der Befrager sollte damit rechnen, daß er den Gesprächsverlauf beeinflusst, ohne es zu wollen, gar ohne es zu bemerken. Mit Gesten, durch Kopfnicken, durch verbale Zustimmung oder Ablehnung kann er ‚Verstärkung‘ oder ‚Bestrafung‘ signalisieren. Er sollte damit rechnen, daß er einzelne Fragen falsch formuliert (sogar damit, daß er ‚aus der Rolle fällt‘). Er sollte wissen, ob er dazu neigt, verletzende Fragen zu stellen, und anerkennen, daß er nicht zu jedem Probanden einen Kontakt herstellen kann, der den Informationsfluß begünstigt.

Gespräche mit Kollegen besprechen: Um mit solchen ‚Schwächen‘ zurechtzukommen, tut der Untersucher gut daran, sie mit Kollegen zu besprechen.

Mit Kollegen sollte er auch, wenn immer möglich, in einer Art Supervision seine Gespräche analysieren. Solche ‚Bearbeitung‘ dient der Erhellung des ‚Falles‘ und der Entdeckung des subjektiven Einflusses, den der Befrager ausübt.

Resümee: Der Abschnitt über die Vorbereitung einer Exploration sei abgeschlossen mit einem Zitat - in Kasten 8-4.

Kasten 8-4: Ratschläge für einen Interviewer

Einige ‚Faustregeln für den Befrager, hat Loretto publiziert (1986, 104).

„*The good interviewer*

- *has a plan,*
- *has adequate job knowledge,*
- *has adequate background information on the applicant,*
- *schedules interviews with adequate time allotted,*
- *ensures that interviews are held in private,*
- *puts the applicant at ease,*
- *lets the applicant talk,*
- *avoids leading questions,*
- *adjusts the level of the language to the ability of the respondent,*
- *is aware of his or her own prejudices and tries to avoid their influence on judgments,*
- *avoids any suggestions of discrimination.*
- *knows how and when to close the interview,*
- *records the facts during the interview and impressions and judgments immediately thereafter“*

8.2.2 Durchführung von Gesprächen

Die Durchführung eines Gesprächs stellt den Untersucher vor neue Anforderungen (Fisseni, Olbrich, Halsig, Mailahn & Ittner, 1993, 224-238).

Die Stichworte könnten lauten:

- dem Befragten Vertrauen entgegenbringen,
- das Gespräch selber steuern,
- das Gespräch gliedern,
- den Befragten ausführlich reden lassen,
- das Gespräch um die relevanten Themen zentrieren,
- ‚Schlüsselbemerkungen‘ des Befragten aufgreifen,
- nonverbale Signale beachten.

Dem Befragten Vertrauen schenken, ihn nicht von vorneherein als ‚Fälscher‘ einstufen

Die Exploration kann Situationen erschließen, so wie das Individuum sie erlebt. Einen Einblick in die Vielfalt des Verhaltens gewährt sie allerdings nur dann, wenn der Befrager dem Befragten Vertrauen schenkt. Wer im Gespräch nur die Gelegenheit sucht, „den Partner beim Lügen oder ‚Aufschneiden‘ zu ertappen,, (Thomae, 1987, 113), wird nicht viel Zutreffendes erfahren.

Das Gespräch selber steuern, die Kontrolle nicht abgeben an den Befragten

Im Gespräch möchte der *Befrager* Informationen *erhalten*. Der *Befragte* möchte Informationen geben, die seinen Vorstellungen entsprechen.

Darum kann der *Befragte* versuchen, selber das Gespräch zu steuern: Er kann beispielsweise

- Sachverhalte verschleiern, die ihn in ungünstigem Licht zeigen,
- Interpretationen anbieten für Sachverhalte, die gegen ihn sprechen,
- ablenken und den Entrüsteten spielen, wenn ‚Ungünstiges‘ zur Sprache kommt.

Was kann der *Befrager* tun, um selber das Gespräch zu kontrollieren? Ein oft erteilter Ratschlag hilft in vielen Fällen weiter: Sobald der Untersucher an Aussagen des Befragten zu zweifeln beginnt, sollte er von ihm verlangen, seine Angaben zu konkretisieren. (*Frage: „Können Sie mir ‚das‘ an einem Beispiel veranschaulichen?“*)

‚Konkrete Verhaltensabläufe‘ zu schildern, wo möglich mehrmals und aus verschiedener Sicht: diese Aufgabe beschränkt die Möglichkeit des Befragten erheblich, mit ‚erfundenen‘ Geschichten aufzuwarten.

Das Gespräch gliedern

Es legt sich nahe, das Gespräch in drei Abschnitte zu gliedern: in Eröffnung, Hauptteil, Abschluß.

Zur ‚**Eröffnung**‘ gehören Begrüßung und Themenangabe. Diese Phase soll dem Befragter den Eintritt ins Gespräch erleichtern, soll seine Ängste mindern.

- *Es empfehlen sich Eisbrecher-Fragen: Wann sind Sie zuhause weggefahren? Wie ist die Anfahrt verlaufen?*
- *Der Befragter kann erklären, wie er sich den Verlauf vorstellt.*

Der **Hauptteil** umfaßt die eigentliche Exploration:

- *Der Befragter soll dem Befragten möglichst rasch das Wort geben und es ihm ‚lange belassen‘.*
- *Die einzelnen Abschnitte der Exploration sollte er als unabhängige Phasen behandeln. Diese ‚Unabhängigkeit‘ soll besagen: Dem Probanden soll gestattet sein, in unterschiedlichen Phasen unterschiedlich ‚exakte Informationen‘ zu geben. Wer in der ersten Phase Sachverhalte beschönigt, muß es nicht auch in der dritten Phase tun.*

Der **Abschluß** sollte dazu dienen, den Befragten zurück ins ‚normale Leben‘ zu entlassen. Der Untersucher kann das Gespräch beenden, indem er den Probanden fragt, ob er seine Angaben ergänzen und modifizieren wolle, ob er das Gefühl habe, daß er alles habe sagen können und fair behandelt worden sei.

Dem Befragten die angemessene ‚Redezeit‘ einräumen

Der Befragte ist der Informant im Gespräch. Darum muß der Befragter ihm ein angemessenes ‚Rederecht‘ und eine ausreichende ‚Redezeit‘ einräumen. Der Befragter läuft Gefahr, zu häufig das Wort zu ergreifen, aber trotzdem die Spanne seiner Redezeit zu unterschätzen.

Eine Orientierungshilfe könnte er finden, wenn er sich an dem sogenannten *Trichtermodell* orientiert: Er läßt den Befragten zu einem Thema zunächst ‚in aller Breite‘ reden. Dann versucht er, ihn zu bewegen, seine Informationen immer enger um den Kern des Themas zu zentrieren.

Diese Technik kann auch dazu beitragen, Angaben zu allen Hauptfragen zu erhalten und den Umfang der ‚fehlenden Daten‘ (missing data) zu verringern.

Das Gespräch zentrieren um die relevanten Themen

Jeder, der Gespräche geführt hat, kennt die Gefahr: Ein Gespräch mündet in eine Sackgasse, weil sich der Befragte vergleichsweise karg äußert und der

Befrager ihm (aus Unsicherheit?) auch dann noch das Wort läßt, wenn der Proband das Hauptfeld verläßt und auf Gebiete überwechselt, die interessant sind, aber nicht zur Sache gehören.

Aufgabe des Untersuchers ist es, den Probanden zum eigentlichen Thema zurückzurufen.

Sensibel sein für Schlüsselbemerkungen des Befragten

Ohne es zu beachten, kann der Befragte - gleichsam nebenbei - ‚seine Situation erklären‘. In einem Nebensatz kann eine Bemerkung fallen, bei der die Stimmlage sich verändert oder Versprecher sich häufen. Solche Hinweise könnten Schlüsselbemerkungen sein. Wie soll der Befrager reagieren?

In jedem Falle sollte er eine Schlüsselbemerkung aufgreifen und sie mit dem Befragten ‚bearbeiten‘.

***Beispiel:** Ein Proband erzählt, daß er Arzt werden möchte, hauptsächlich um anderen Menschen zu helfen; dann sagt er: „Helfen kann man ja auch in anderen Berufen. Aber meine Mutter will mein Vater war ja auch Arzt, der ist seit sieben Jahren tot, ein hier in der Stadt bekannter Gynäkologe war er, und meine Mutter denkt, daß ich von seinem Rufe einmal ganz schön profitieren kann... Ich meine, von mir aus...! Jurist wäre aber auch ganz schön...“*

Die ‚Abschweifungen‘ des Probanden könnten ‚verraten‘, daß der Berufswunsch ‚Arzt zu werden‘, extern gesteuert ist (eingegeben von der Mutter), daß der interne Berufswunsch eher in Richtung eines Jurastudiums geht. - Für die Berufswahl könnte es entscheidend sein, daß der Untersucher dieses Problem anspricht und versucht, es mit dem Probanden zu klären.

Nonverbale Signale beachten

Ein Informationsquelle eigener Art kann das nonverbale Verhalten sein. Dazu gehören Mimik, Gestik, Körperhaltung, Sprachmodulation.

- Zwischen *Körperhaltung* und Erleben besteht eine Wechselwirkung. Das Erleben drückt sich in der Körperhaltung aus, die Körperhaltung ‚formt‘ das Erleben. (*Warum setzt sich der Proband eigenartig hin: vorgebeugt mit Blick auf den Boden?*)
- Auffällig kann der *Wechsel* einer Körperhaltung sein (*etwa, wenn sich der Proband plötzlich vom Befrager wendet*).
- Überraschend kann auch eine *Änderung* der Sprachmodulation wirken (*etwa, wenn der Proband plötzlich in hastiges Sprechen verfallt*).
- Eine *plötzliche Gesprächspause* kann den Untersucher befremden, ja beunruhigen. (*Ohne ersichtlichen Grund verstummt der Proband.*)

Nonverbale Signale können zusätzliche Informationen liefern - *aber welche Informationen?* Das Problem besteht darin, daß keineswegs eindeutig ist, was eine bestimmte Körperhaltung oder ein bestimmter Tonfall ausdrücken. Ihre Auswertung birgt darum jederzeit die Gefahr einer Fehldeutung in sich, beispielsweise einer Überinterpretation.

Statt vorrangig nach inhaltlichen Deutungen zu suchen, könnte der Untersucher auch anders vorgehen; er könnte nonverbale Signale als Zeichen betrachten, in denen sich *die aktuelle emotionale Lage* des Probanden ‚äußert‘.

Bei dieser Sicht erschiene es sinnvoll, die ‚Zeichen‘ mit dem Probanden zu besprechen, zu ‚bearbeiten‘ und - unter seiner Mithilfe - gegebenenfalls auch zu deuten. (*Ich habe den Eindruck, daß unser Gespräch Sie mitnimmt. Ist es Ihnen recht, wenn wir jetzt einmal über diesen Punkt reden? Können Sie mir sagen, was Sie gerade jetzt bewegt?*)

Resümee: Der Befrager muß darauf achten, daß er die Kontrolle über das Gespräch wahr, sich aber ebenso dringlich um eine Atmosphäre bemüht, in welcher der Befragte sich frei äußern und ‚sich erklären‘ kann.

8.2.3 Zur Auswertung von Gesprächen in der Diagnostik

Standardisierte Interviews liegen nach der Kodierung ausgewertet vor. Die Aufgabe einer besonderen Auswertung stellt sich nur für *halb- und unstandardisierte Gespräche*.

Hinweise zu Auswertung geben viele Autoren, beispielsweise Fisseni, Olbrich, Halsig, Mailahn & Ittner 1993, 239-245; Jahoda, Deutsch & Cook, 1965, 271-289; Noelle, 1968, 205-218; Undeutsch, 1983, 334-336; Westhoff & Kluck, 191, 164-167; 214; Wittkowski, 1994, 15-21; Zeisel, 1965, 290-318.

Formal muß eine Auswertung gewährleisten, daß sie

- die relevanten Teile des Gespräches *vollständig* wiedergibt,
- die Inhalte *eindeutig* darstellt und
- die Aussagen, so weit möglich, in *disjunkten Kategorien* zusammenfaßt (von Hagen, 1988, 226).

Inhaltlich muß eine Auswertung sich an zwei Gegebenheiten orientieren:

- Die **Art** der **Registrierung** begrenzt oder erweitert die Möglichkeiten einer Auswertung: Wird ein Gespräch auf Band oder Video mitgeschnitten, kann seine Auswertung detaillierter geplant werden, als wenn eine Mit- oder Nachschrift vorliegt.
- Das **Ziel** legt den Umfang und die Art der Wiedergabe fest: Ist eine Auswertung für ein *Gutachten* vorgesehen, muß der Gutachter die Inhalte fallbezogen wiedergeben, so eindeutig und vollständig, daß sie in der Stel-

lungnahme diskutiert werden können. - Dient eine Auswertung einer Beratung, genügt es, wenn sich der Berater mit den Inhalten vertraut macht und aus ihnen die Hypothesen entwickelt, die sein weiteres Vorgehen leiten sollen.

Skizziert seien *zwei Grundmodelle* der Auswertung:

- Wiedergabe des Originalgespräches (8.2.3.1),
- Zusammenfassung des Gespräches (8.2.3.2).

8.2.3.1 Wiedergabe des Originalgespräches

Die Exploration wird auf Video oder Tonband festgehalten, kann auf Wunsch reproduziert, gegebenenfalls auch vollständig transkribiert werden.

Diese Prozedur kommt für diagnostische Fragestellungen in der *Regel* kaum in Betracht. Der Aufwand wäre hoch, der Nutzen zu gering - gering, weil das Gespräch noch gar nicht ‚diagnostisch bearbeitet‘ worden ist.

Doch kann sich die Wiedergabe von Originalen mindestens in drei Fällen als **nützlich erweisen**:

1. Beim *Training in Gesprächsführung* lassen sich anhand von Video- oder Tonbändern gelungene und fehlerhafte Passagen vorspielen.
2. Während einer Beratung können dem Probanden zur *Demonstration* Originalszenen vorgespielt werden.
3. Wird ein Gutachten mündlich vorgetragen, etwa vor Gericht, dann können ausgewählte Abschnitte der originalen Aufnahme die *emotional-affektive* Tönung von *Aussagen* veranschaulichen.

8.2.3.2 Zusammenfassung eines Gespräches

Zwei Arten von Auswertung seien skizziert:

- Schematische Zusammenfassung (A),
- Thematische Zusammenfassung (B).

(A) Schematische Zusammenfassung

Für Aufgaben, die keinen ausformulierten Bericht erfordern, läßt sich die Exploration schematisch zusammenfassen.

- Konzipiert oder übernommen werden **verbale Schemata**, in die das Gespräch übertragen wird. Kasten 8-5 bringt ein Beispiel.

Kasten 8-5:
Beispiel für ein verbales Auswertungsschema

Quelle: „Diagnostischer Elternfragebogen (DEF)“ von Dehmelt, Kuhnert und Zinn (1981)

V. BEZIEHUNGEN ZU ANDEREN PERSONEN		
Auswerten: Spalte 2 ankreuzen!	2	Verhaltensweise
41. Wie verhält sich Ihr Kind in Gruppen außerhalb der Schule?		<i>kommt gut mit den anderen aus</i>
		<i>sucht andere zu beherrschen</i>
		<i>leicht durch andere zu beeinflussen</i>
		<i>will häufig alles bestimmen, rechthaberisch</i>
		<i>nimmt anderen gerne etwas weg</i>
		<i>zwickt, stößt, schlägt andere häufig</i>
		<i>wird von anderen häufig gezwickt, gestoßen geschlagen</i>

- Entworfen oder übernommen werden **numerische Schemata**, verbunden mit verbalen Ankern, die den Zahlen ‚Bedeutungen‘ zuweisen. Kasten 8-6 bringt ein Beispiel.

Kasten 8-6:
Beispiel für ein numerisches Auswertungsschema

Quelle: „Beobachtungsbogen für Kinder im Vorschulalter (BBK 4-6)“ von Duhm und Althaus (1979)

Abschnitt II: <i>Soziales und emotionales Verhalten:</i>					
<i>Auswertung: Fünf Skalenwerte geben an, wie häufig ein Verhalten auftritt.</i>	1 Sehr selten	2 Manchmal	3 Teils, teils	4 Oft	5 Sehr oft
7. <i>Das Kind erzählt und berichtet von sich aus anderen Kindern.</i>					
8. <i>Es erzählt und berichtet von sich aus der Erzieherin.</i>					
9. <i>Es stellt Fragen und will viel wissen.</i>					
10. <i>Es äußert seine Wünsche.</i>					
11. <i>Es macht Vorschläge.</i>					
11. <i>Es nimmt aktiv am Gruppen-geschehen teil.</i>					

Zwischen verbalen und numerischen Schemata läßt sich eine **Vielfalt unterschiedlicher Mischformen** bilden.

(B) Thematische Zusammenfassung

Für Aufgaben, die eine ausführliche (aber keine wörtliche) Wiedergabe des Gespräches erfordern, sei eine Sonderform der Auswertung skizziert - die ‚Thematische Zusammenfassung‘. Sie setzt voraus, daß ein Gespräch vollständig aufgezeichnet worden ist (z.B. auf Tonband oder Video). Vorgeschlagen seien drei Schritte (Fisseni, 1992, 164):

- **Schritt 1: Themenbereiche identifizieren**

Der Auswerter geht die Aufzeichnung durch, um relevante Themen zu identifizieren, *etwa Interessen, Beziehung zu den Eltern, Leistungsmotivation, Tendenzen zur Anpassung usw.*

Textstellen, die zu einem Thema Aussagen enthalten, werden *markiert*: Geschriebene Texte können unterstrichen, abgehörte Texte mit ihrer Nummer notiert werden.

- **Schritt 2: Aussagen zu Themenbereichen zusammenstellen**

Aus den markierten Stichwörtern sammelt der Auswerter die Belege und notiert sie auf *Einzelblättern*, je Thema ein Blatt (oder auf dem PC je Thema eine kurze Datei).

Um die Übersicht zu wahren, sollte er eine *knappe, geraffte Darstellung* anstreben.

- **Schritt 3: Verarbeitung zu einem fortlaufenden Text**

Die Themenbereiche (auf den einzelnen Blättern) werden in eine Reihenfolge gebracht, die der diagnostischen Fragestellung angemessen ist.

Dann wird aus ihnen *einfortlaufender Text* erstellt, gegliedert nicht mehr nach dem Verlauf der Exploration, sondern nach den Gesichtspunkten, die dem Diagnostiker sinnvoll erscheinen.

Vorschläge zur *sprachlichen Gestaltung* referiert Kasten 8-7.

Eine Thematische Auswertung anzufertigen erfordert einen hohen Aufwand; er lohnt sich nur, wenn er erheblich zur Klärung der diagnostischen Frage beiträgt, zum Beispiel im Rahmen einer umfassenden Begutachtung.

Kasten 8-7:

Zur sprachlichen Gestaltung einer Thematischen Zusammenfassung

Die Thematische Zusammenfassung gibt die Schilderung des Probanden genau, *aber gekürzt* wieder. (Dabei sollte die Wiedergabe die Sicht des Probanden beibehalten, sie sollte seine subjektive Schilderung zu erkennen geben.)

Der *Spruchduktus des Probanden* sollte, so weit möglich, erkennbar bleiben, etwa seine Wortwahl oder sein Satzbau.

Am Beginn kann der Auswerter das *Verhalten* des Probanden kennzeichnen (z.B. „Er ging auf Fragen nicht ein.“ - „Er mied jeden Blickkontakt.“)

Zur Kennzeichnung der mittelbaren Wiedergabe sollte man die *Aussagen im Konjunktiv* referieren (z.B. „Er berichtete, er *habe* seinen Bruder nie gemocht.“)

Fakten, unstrittige Angaben erscheinen *im Indikativ* (Alter, Beruf, Namen usw.).

Für die Gliederung empfiehlt es sich, biographische Angaben *chronologisch* zu berichten, dagegen Aussagen über die augenblickliche Situation *thematisch* zu ordnen.

Im Text setzt man *keine Überschriften*, doch können Stichworte *unterstrichen* oder *gesperrt* werden.

Zitate sollte man *sparsam* verwenden. Sie tauchen nur auf, wenn sie besonders *charakteristisch* (und für den Probanden nicht verletzend) sind.

Eine **Interpretation** der Exploration ist ein zusätzlicher Auswertungsschritt; sie kann als Zusammenfassung am Schluß erscheinen.

Kasten 8-8 referiert ein Beispiel für die Thematische Zusammenfassung.

Kasten 8-8:

Beispiel einer Thematischen Zusammenfassung

Das Gespräch, aus dem wir Auszüge bringen, wurde auf Band mitgeschnitten und wörtlich transkribiert. Untersucher war ein Psychologe (in einer Erziehungsberatungsstelle). Proband war Karsten (14 Jahre alt). Problem: Vater und Sohn „kommen nicht miteinander aus“. Die Familie besteht aus Vater, Mutter und zwei Jungen: Karsten und Arno (9 Jahre).

Schritt 1:

Themenhereiche identifizieren

HINWEIS: Es folgen die Abschnitte (1) bis (6) aus der Exploration. Das Originalgespräch umfaßte 44 Abschnitte. Es handelte sich um eine unstandardisierte Exploration

- ihre Schwächen werden erkennbar: auch wenn wir nur kurze Auszüge zitieren
- Identifiziert wurden sieben Themen: Vater / Mutter / Bruder / Schul- und Leistungsverhalten / Kontakte außerhalb der Familie / Selbsteinschätzung / Freizeit.
- In Klammern stehen die Fragen des Untersuchers.

Bei dieser Demonstration beschränken wir auf die Identifizierung zweier Themen: (1) Verhältnis zum Vater, (2) Verhältnis zur Mutter:

- Die zwei ersten Aussagen, die sich auf den **Vater** beziehen, unterstreichen wir
- Die zwei ersten Aussagen, die sich auf die **Mutter** beziehen, setzen wir kursiv.

(1) (Karsten, Dein Vater war bei mir und hat mir von Eurer Familie erzählt.) Ich weiß, er hat wohl etwas gesagt, hm, ja, daß er sich früher falsch verhalten hat, in der Erziehung, ja, daß ... er hat uns falsch erzogen.

Wieso?) Ja, ich weiß nicht, was er gesagt hat, aber jedenfalls mit den falschen Methoden, er hat uns falsch verstanden oder sich nicht darüber informiert oder was. Er hat auch schon mit uns darüber gesprochen.

Und was ist falsch gelaufen?) Ja, daß er überarbeitet war, daß er manchmal aus der Haut geraten ist, daß er geprügelt hat. (Geprügelt?) Ja, auch, das ist aber schon lange her. Aber ich meine, er fühlt sich jetzt irgendwie belastet. Vielleicht sieht er bei uns irgendwelche Fehler und fuhr die auf seine Erziehung zurück.

(2) (Was meinst Du mit den Fehlern?) Ja, ich könnte mir denken, daß ich irgendwie zu still bin. Ich meine, ich bin mir zwar bewußt, daß ich etwas still bin, aber bei mir gibt's einfach nicht so viel zu reden. Also, also erst einmal, ich führe nicht so gerne Gespräche über Themen, die mich nicht interessieren, also was ich am liebsten mache am Tag und so. Ich unterhalte mich lieber über Dinge, die für mich interessant sind.

(3) (Was ist interessant?) Ja, zum Beispiel meine Hobbies. Ja, ich interessiere mich für Radio-technik, dann spiele ich Gitarre und Klavier, wir haben zwei Hunde, mit denen beschäftige ich mich auch. Und auch Kakteen! Naja, jetzt nicht, die haben jetzt ihre Ruhezeit, da kann man nicht viel machen.

(4) (Was interessiert Dich im Augenblick?) Radiotechnik, wie man eben so ein Radio baut. Also nicht Funktechnik, nicht Fernsehen, sondern Radiotechnik, wie man so ein Radio baut. Ich habe einen Detektor gebaut, ich hab so von Kosmos einen Experimentierkasten. Als nächstes bau ich einen Transistorverstärker. (Mit wieviel Watt?) Ich hab den noch nicht im Augenblick, das steht da in den Plänen drin, ich hab so ein Baubuch, habe ich aber noch nicht angeguckt.

(5) (Könntest Du mir bitte erzählen, wie ein Tag bei Dir verläuft, zum Beispiel gestern: Wie ist es gestern gelaufen?) Ja, ich hab gestern den ganzen Tag im Bett gelegen und hab mittags mit meinem Bruder gespielt. Wach geworden bin ich, glaub ich, um 9 Uhr. Ich hab dann noch etwas geschlafen, dann habe ich Felix-Hefte gelesen. Dazwischen habe ich noch gefrühstückt? (Wo?) Im Bett! *Meine Mutter hat auf dem Tablett alles gebracht. Dann..., ich hab ihr gesagt, wie ich mich fühle, ja, das hat sich so ergeben, sie hat mir noch ein paar Halstabletten gegeben.*

(6) (Wie verstehst Du Dich mit Deiner Mutter?) *Ja, so ganz gut! Aber, ja, ich weiß nicht, was Ihre Frage..., die Frage ist jetzt..., da kann man sich viel drunter vorstellen.*

(Wenn Du ein Problem hättest, gingst Du zu deiner Mutter?) Nein, eher würd ich mich mit meinem Vater unterhalten, ja, ich glaube, eher mit dem Vater. Mit der Mutter auch, hm, ja, aber ... nicht zuerst.

(Auch mit einem Freund?) Ne, mit ,nem Freund nicht. Ich hab ja da auch keinen, zu dem ich gehen kann. Wo ich in die Schule geh, da ist es schwer, bei mir in der Gegend wohnt keiner.

Schritt 2:

Aussagen zu Themenhereichen zusammenstellen

HINWEIS: In unserem Beispiel wurden sieben Themen identifiziert, somit lägen am Ende sieben Blätter vor. - Unsere Demonstration bringt nur Stellen zu den Themen „Vater“ und „Mutter“ aus den Abschnitten 1 bis 6.

Thema „Vater“

Abschnitt

ja, daß er sich früher falsch verhalten hat, in der Erziehung, ja, daß . . . er hat uns falsch erzogen. 1

Ja, daß er überarbeitet war, daß er manchmal aus der Haut geraten ist, daß er geprügelt hat. 1

er fühlt sich jetzt irgendwie belastet. Vielleicht sieht er bei uns irgendwelche Fehler und führt die auf seine Erziehung zurück. 1

(Wenn Du ein Problem hättest, gingst Du zu Deiner Mutter?) Nein, eher würd ich mich mit meinem Vater unterhalten, ja, ich glaube, eher mit dem Vater. 6

Thema „Mutter“

Abschnitt

Meine Mutter hat (mir ans Bett zum Frühstück) auf dem Tablett alles gebracht. Dann..., ich hab ihr gesagt, wie ich mich fühle, ja, das hat sich so ergeben, sie hat mir noch ein paar Halstabletten gegeben. 5

(Wie verstehst Du Dich mit Deiner Mutter?) Ja, so ganz gut! Aber, ja, ich weiß nicht, was Ihre Frage..., die Frage ist jetzt..., da kann man sich viel drunter vorstellen. 6

(Gingst Du mit einem Problem zu Deiner Mutter?) Nein, eher würd ich mich mit meinem Vater unterhalten, ja, ich glaube, eher mit dem Vater. Mit der Mutter auch, hm, ja, aber . . . nicht zuerst. 6

Schritt 3: Verarbeitung zu einem fortlaufenden Text

HINWEIS: Beim dritten Schritt werden die Themen zu einem fortlaufenden Text **verarbeitet**. Es ist die schwierigste und aufwendigste Aufgabe.

- Die „Thematische Zusammenfassung“ in diesem Demonstrationsbeispiel bezieht sich nur auf „Vater“ und „Mutter“, soweit Aussagen in den sechs zitierten Abschnitten vorkommen.
- Dagegen bezieht sich die „Zusammenfassende Interpretation“ auf die gesamte Exploration.

Mit Karsten M. wurde am 12. Mai 1995 ein Gespräch geführt, das etwa sieben Minuten dauerte. Häufig antwortete Karsten nur stockend und mußte zu weiteren Antworten ermuntert werden.

Im Gespräch kamen folgende **Themen** zur Sprache: die Beziehung zu seinem Vater, zu seiner Mutter und zu seinem Bruder, die Kontakte zu Gleichaltrigen, der Leistungsbereich vor allem im Rahmen der Schule, die Hobbies und die Art seiner Selbsteinschätzung.

Über das Verhältnis zu seinem Vater machte Karsten gegensätzliche Angaben. Zum einen sagte er, der Vater habe ihn und seinen Bruder Arno früher mit falschen Methoden erzogen; vermutlich habe er seine beiden Söhne falsch verstanden oder sich ungenügend über eine richtige Erziehung informiert. Zum anderen bezeichnete Karsten den Vater als Person seines Vertrauens mit dem er über seine Probleme reden könne; der Vater seinerseits habe auch mit ihm (Karsten) und seinem Bruder (Arno) über Erziehungsfragen gesprochen und zugegeben, daß er sich früher falsch verhalten habe.

Das frühere Verhalten des Vaters (Prügelstrafe) entschuldigte Karsten geradezu, indem er sagte, der Vater sei wohl überarbeitet gewesen, wenn er aus der Haut geraten sei und geprügelt habe. Von diesem Mißgriff fühle sich der Vater auch heute noch belastet - vermutlich, weil er bei seinen Kindern Fehler entdeckte, die er auf seine falsche Erziehung zurückführe. Usw.

Über seine **Mutter** redete Karsten vergleichsweise wenig. (Vielleicht hat der Psychologe zu wenig nachgefragt!) Karsten betonte aber, daß er sich ganz gut mit ihr verstehe. Er stellte einen fürsorglichen Zug der Mutter heraus, weil sie ihn geradezu verwöhnend versorgt habe, als er krank im Bett lag. Als Vertrauensperson steht sie hinter dem Vater zurück: Wenn er Probleme habe, wende er sich zunächst an den Vater, erst später auch an die Mutter. Usw.

Zusammenfassende Interpretation

Das Ergebnis der gesamten Exploration läßt sich wie folgt zusammenfassen:

Der **Vater** steht für Karsten als Vertrauensperson an erster Stelle. Er ist Ansprechpartner, wenn Schwierigkeiten zu besprechen sind; er gilt als Kollege und Freund, wenn es um praktische Tätigkeiten geht - der Vater verstehe viel von Basteln. In der Vergangenheit sei das Verhältnis eine Zeitlang sehr gespannt gewesen, weil der Vater eine „falsche Erziehung“ praktiziert habe.

Über die **Mutter** äußert Karsten sich karg und eher vage. Ein gewisse Bedeutung gibt er dem Fürsorgeaspekt. Ansprechpartnerin für Probleme ist sie erst in zweiter Linie.

Der **Bruder** ist für Karsten einerseits ein wichtiger Spielgefährte, andererseits kommt es oft zu Reibereien und Rivalitäten zwischen ihnen.

Die **Schule** beurteilt Karsten insgesamt mit Zustimmung. Was die Leistungen angeht, so betonte er, daß er sich den Anforderungen stelle (er müsse allerdings zuhause viel für die Schule arbeiten); er halte sich aber nur für einen ‚mittelguten‘ Schüler. Was die **soziale Anerkennung** betrifft, so erzählte er, daß er die Achtung seiner Kameraden erst nach und nach gewonnen habe, sich aber heute von der Klasse akzeptiert fühle.

Nur wenige Kontakte unterhält Karsten zu **Gleichaltrigen**, er äußert sich distanziert über frühere Bekannte, die in der Hauptschule geblieben seien. Zu einem Klassenkameraden hat er ein engeres freundschaftliches Verhältnis; aber so eng sei die Beziehung nicht, daß er mit diesem Freund auch persönliche Probleme besprechen könne.

Was **Freizeit und Hobbies** angeht, so beschäftigt er sich am liebsten mit Tätigkeiten, die es ihm ermöglichen, allein zu sein: Er höre gerne Musik, befasse sich mit technischen Aufgaben, vor allem mit Radiotechnik. Ein wichtiger ‚Partner‘ ist für ihn sein Hund. Ein anderer Gegenstand seiner Sorge und Aufmerksamkeit sind seine Pflanzen - vor allem Kakteen.

Er erlebt **sich selbst** als zurückgezogen und beurteilt seine soziale Umwelt vergleichsweise kritisch und bekundet dabei ein hohes Zutrauen in sein eigenes Urteil.

8.2.4 Beitrag zu Diagnostik und Intervention

Was leistet das Gespräch für die Diagnostik und was für die Intervention?
Diese Frage sei in zwei Abschnitten beantwortet:

- Beitrag für die Diagnostik,
- Beitrag für die Intervention.

Beitrag explorativer Techniken zur Diagnostik

In der diagnostischen Forschung oder bei Reihenuntersuchungen, vor allem aber in der Individualdiagnostik, empfiehlt sich die Exploration an verschiedenen Stellen:

1. Die wichtigste Rolle spielt sie bei der Erarbeitung der diagnostischen Fragestellung („Warum sind Sie zu uns gekommen?“). In der Individualdiagnostik sollte die Exploration in halb- oder unstandardisierter Form geführt werden. Wenn dieses Gespräch mißlingt, ist der Mißerfolg für die Gesamtuntersuchung vorgezeichnet.
2. Kaum zu ersetzen ist die Exploration auch
 - bei *Abklärung des Kontextes*, in dem die zentrale Frage steht, so etwa, wenn es um das Umfeld ‚Schule‘ oder ‚Beruf‘ geht (Klima, Zufriedenheit, Beziehung zu Kollegen, Berufsperspektive);
 - ebenso bei *Erhebung persönlicher und intimer Informationen*, so etwa, wenn die Lebensgeschichte, die Beziehung zu Partnern (Gatten, Geschwistern, Freunden, Bekannten, Kollegen), die Interessensstruktur, die beruflichen oder finanziellen Perspektiven ermittelt werden sollen.
3. Eine ähnliche Funktion fällt der Exploration zu
 - bei dem Vergleich von *Informationen* mit dem Ziel, sie zu einer Stellungnahme zu verarbeiten;
 - bei der *Schlußmitteilung* an den Probanden, vor allem bei Beratung oder Zuweisung zu einer Intervention.

In diesen Beispielen, vor allem wieder bezogen auf **Einzelfalldiagnostik**, kann eine Befragung mittels standardisierter Leitfaden das ‚offenere‘ Verfahren der Exploration vorbereiten, aber nicht ersetzen.

Beitrag explorativer Techniken zur Intervention

Für alle Varianten, in denen Intervention auftritt, können Interview, Anamnese, Exploration Dienste übernehmen. Doch sind sie im interventiven Kontext nie isoliert zu betrachten, sondern immer zu sehen im Verbund mit anderen Instrumenten.

In unserer Aufzählung wird allerdings der spezielle Anteil explorativer Techniken wie unter einer Lupe vergrößert.

Wir bringen Beispiele aus zwei Bereichen:

(1) Drei Funktionen kann das Gespräch in den verschiedenen *Therapieschulen* übernehmen: Es kann Störungen identifizieren (helfen), es kann als Korrekturinstrument dienen, es kann das Ergebnis evaluieren (helfen).

Vorausgeschickt sei eine Bewertung von Scholz - es geht um die Bedeutung, die das „Interview“ im Dienste diagnostisch-interventiver Maßnahmen bei Partnerschaftstörungen erhält. Die Bewertung läßt sich auf ähnliche diagnostisch-therapeutische Situationen übertragen.

„Das diagnostische Gespräch dürfte nach wie vor als eine unverzichtbare Methode für die Bewertung ehelicher Störungen anzusehen sein. Das Anliegen des Interviewers, Informationen über die Dyade zu sammeln, sollte dann verfolgt werden, wenn geeignetere diagnostische Methoden nicht verfügbar sind. Interviews sollten aber auch mit gleicher Wichtigkeit geführt werden, um damit wesentliche Vorbedingungen für einen psychotherapeutischen Erfolg zu schaffen: Dies kann beispielsweise in der Absicht geschehen, daß im gemeinsamen Gespräch positive Erwartungen und Mut, das Angebot des Psychologen anzunehmen und aktiv mitzumachen, bei beiden Partnern erzeugt werden. Traditionell wird mit dem Interview das Anliegen verfolgt, eine Spezifizierung der Zielvorstellungen der Intervention einschließlich deren Antezedenzen und Konsequenzen zu erreichen und darüber hinaus erste Hinweise für die Formulierung einer Interventionsstrategie zu erreichen.“

Quelle: „Ehe- und Partnerschaftstörungen“ von Scholz (1987, 84)

Nur *angespielt* sei auf Psychoanalyse, Gesprächspsychotherapie, Verhaltens- und systemische Familientherapie.

Psychoanalyse: Aufgedeckt werden Störungs-Ursachen in einer speziellen Art verbaler Bewußtmachung. Aus Träumen, die der Klient erzählt und unter Assistenz des Therapeuten ‚deutet‘, aus freien Assoziationen, die der Klient preisgibt und unter ‚Anleitung‘ des Therapeuten interpretiert, wird die Vergangenheit als traumatisierender Grund rekonstruiert. - Auch korrigiert werden die Störungen in verbaler Interaktion, sofern in den wiederbelebten Erinnerungen die Vergangenheit neu durchlebt, neu durchlitten und in eine annehmbare Ge-

stalt umgewandelt wird. - Ob das therapeutische Ziel erreicht ist, ob also die Genuß- und Leistungsfähigkeit wiederhergestellt ist, wird wiederum erkennbar aus ‚Gesprächen‘.

Gesprächspsychotherapie: Das Gespräch dient der Enthüllung einer Inkongruenz in den Gefühlen und Wünschen des Klienten. Eine solche Inkongruenz wird angenommen, wenn der Proband es nicht ‚wagte‘, die authentischen Gefühle und Wünsche seines wahren Selbst zu äußern und ihnen zu folgen, sondern sie an die Forderungen wichtiger Bezugspersonen (etwa der Eltern) angepaßt, sie umgeformt und dabei auch deformiert hat. - Im Dialog mit dem Therapeuten soll der Klient lernen, seine ‚wahren‘ Gefühle zu ‚bemerken‘ und zuzulassen, sie zu äußern, zu billigen und zu akzeptieren (Dahmer & Dahmer, 1982). - In diesem Dialog übt der Klient die Fertigkeit ein, seine ursprünglichen Strebungen und Kompetenzen zu gebrauchen und zu verwalten. In einer Autonomie der ‚befreiten‘ Gefühlswelt sollte sich der Erfolg des therapeutischen Prozesses dann manifestieren.

Verhaltenstherapie: Auch in der Verhaltenstherapie (von ihrem Ansatz sicher abhold dem introspektiven Anteil eines Gespräches) spielt das Interview eine entscheidende Rolle (Lutz, 1978). Freilich gilt vor allem in diesem Beispiel, daß ein ‚Interview‘ eingebunden bleibt in ein Repertoire anderer Methoden, etwa der Verhaltensbeobachtung oder spezieller Fragebogen (Schulte, 1976). Aber auch mithilfe von Interviews, werden klar umschriebene Verhaltensweisen (Kognitionen, Emotionen, Motive) identifiziert, die störend wirken und darum verändert werden sollen. - Gespräche begleiten die Konzeption und den Entwurf von Strategien, die es ermöglichen sollen, die gestörten Verhaltenssequenzen durch Umlernen zu korrigieren. Gespräche können die individuelle Passung solcher Strategien kenntlich machen. - Neben anderen Methoden ist es auch das Interview, das dazu beiträgt, den Verlauf und den Erfolg einer Änderungsprozedur zu bewerten.

Systemische Familientherapie: An einem Beispiel der systemischen Familientherapie veranschaulicht Tömm die Funktion zirkulärer Fragen (1994, 182). Ein Therapeut spricht mit Vater, Mutter und den zwei Kindern der Familie Uvland:

- Therapeut: „Frau Uvland, wie kommt es, daß wir uns heute treffen?“
 Frau: „Ich habe Sie angerufen, weil ich mir wegen der Depressionen meines Mannes Sorgen mache.“
 Therapeut: „Wer macht sich sonst noch Gedanken?“
 Frau: „Unsere Kinder!“
 Therapeut: „Wer macht sich Ihrer Meinung nach die größten Sorgen?“
 Frau: „Ich.“
 Therapeut: „Wer sorgt sich Ihrer Meinung am wenigsten deswegen?“
 Mann: „Ich nehme an: Das bin ich.“
 Therapeut (an den Mann): „Wenn Ihre Frau sich sorgt, was tut sie dann?“

- Mann: „*Sie beklagt sich sehr oft, hauptsächlich wegen des Geldes und wegen irgendwelcher Rechnungen.*“
- Therapeut: „*Wie verhalten Sie sich, wenn sie Ihnen zeigt, daß sie sich Sorgen macht?*“
- Mann: „*Ich belästige sie nicht weiter; sondern halte mich zurück.*“
- Therapeut: „*Wer erlebt die Sorgen Ihrer Frau am stärksten?*“
- Ungefragt reden die Kinder dazwischen:
 „*Die größten Sorgen macht sich nicht der Vater auch nicht die Mutter Wer sich die größten Sorgen macht, sind wir Kinder.*“
- Therapeut: „*Seid ihr Kinder wirklich dieser Meinung?*“
- Kinder: „*Ja!*“
- Therapeut: „*Was macht Euer Vater gewöhnlich, wenn Ihr Euch mit eurer Mutter unterhaltet?*“
- Kinder: „*Er geht dann ins Bett.*“
- Therapeut: „*Und wenn er ins Bett geht, was macht dann Eure Mutter?*“
- Kinder: „*Sie macht sich dann noch mehr Sorgen.*“
- usw.

(2) In der **Arbeits- und Organisationspsychologie** können Anamnese, Exploration oder Interview dazu beitragen, Aufgaben der *Personalentwicklung* zu lösen – nur dies eine Beispiel sei skizziert (Becker & Langosch, 1990).

Personalentwicklung beschäftigt sich mit der systematischen Förderung beruflicher Qualifikation und ihrer Einordnung in den Entwicklungsplan einer Organisation. Erforderlich sind demnach unter anderem Eignungsbeurteilungen (Personal) und Vergleiche mit dem Bedarf (Organisation) und Ausbildung der „entdeckten“ Ressourcen (Personal-Entwicklung).

Noch einmal anders gesagt: Personalentwicklung soll eine Verbindung herstellen

- zwischen Unternehmen mit seinen Plänen und Ressourcen sowie
- den Mitarbeitern mit ihren Karrierevorstellungen und Entwicklungspotentialen.

Hier kann sich das Interview, flexibel eingesetzt, als höchst aufschlußreich erweisen. Interviews können helfen, Bedingungen zu erkennen

- für eine Entwicklung individueller Potentiale,
- für eine Steigerung von Gruppeneffektivität,
- für eine Verbesserung der Interessensbeziehungen zwischen den Mitarbeitern und
- für die Integration ihrer Kompetenzen in die Gesamtorganisation.

Es gilt sogar: Je individueller die Eignungsfrage gestellt wird, erst recht aber: je idiosynkratischer eine Stelle ist, für deren Besetzung das Potential abgeschätzt werden soll, um so aufschlußreicher sind die Informationen, welche ein Interview erbringen kann.

8.2.5 Fehlertendenzen

Bei Gesprächen können ähnliche Fehler auftreten wie bei der Verhaltensbeobachtung (vgl. S. 199):

Die **allgemeinen Fehler**, die bei der Verhaltensbeobachtung besprochen wurden (S. 199), können auch eine Exploration beeinflussen, etwa die „zentrale Tendenz“, der „Hof-“ und „Positionseffekt“. Auf diese Fehler sei nur verwiesen.

Für **spezielle Fehler**, die eine Exploration verfälschen können, nennt die Literatur verschiedene Beispiele (Anger, 1969, 598; Atteslander & Kneubühler, 1975; Lutz, 1978, 116-122; Maccoby & Maccoby, 1965, 74-76; Schraml, 1964, 888; 1975; Westhoff & Kluck, 1991, 139-149):

- *Unterschiedliche Frageformulierungen* können Antworten provozieren, die nicht vergleichbar sind, eine Gefahr vor allem des unstandardisierten Gesprächs.
- *Voreinstellungen* bei dem Probanden oder dem Befrager können die Ausagenrichtung lenken (aufgrund unkontrollierter Belohnung oder Bestrafung).
- Die *Dynamik eines Gesprächs* kann in Befragtem und Befrager bestimmte Antworttendenzen erzeugen: Es werden etwa Antworten ‚erwartet‘ (dann auch gegeben und protokolliert), die zu schon geäußerten Aussagen ‚passen‘.
- *Nonverbale Äußerungen* des Befragers (etwa Nicken, Lächeln, Räuspern) können erlebt werden als ‚Belohnung‘ oder ‚Bestrafung‘ und so dem Probanden bestimmte Antworten ‚entlocken‘ oder ‚verbieten‘.
- Das Gespräch kann *selektiv* registriert oder selektiv ausgewertet werden.

8.3 Zu den Gütekriterien explorativer Daten

Die Frage nach Objektivität, Reliabilität und Validität explorativer Daten ist ein heiß umstrittenes Thema. Schulmeinungen beeinflussen die Stellungnahmen.

Eine Durchsicht von mehr als hundert Studien belegt, daß die Ergebnisse sehr differenziert zu bewerten sind (Frinken, 1980; Gielen & Kaden, 1977; Küpper 1993; Schmidt, K. J., 1980; Seek, 1982; Westhofen, 1991).

Die Grundfrage lautet: Läßt sich das Meßmodell der klassischen Testtheorie auf explorative Daten anwenden? Oder umgekehrt gefragt: Werden explorative Daten dem Meßmodell der klassischen Testtheorie gerecht?

Es ist die Frage nach der *Modellverträglichkeit* (Algera, 1976, 3; Cannel & Fowler, 1967, 250; Fisseni, 1974, 32-34; Helzer et al., 1977, 129-130; Lehr, 1964, 98; Lutz, 1978, 20-24; Parry & Crossley, 1950, 69; Schmidt & Keßler, 1976, 12, 96-98; Seidenstücker, 1974, 378-380; Sines, 1959, 483).

Als ein Resümee sei in Kasten 8-9 der Abschluß der einschlägigen Diskussion bei Lutz zitiert (1978, 23-24).

**Kasten 8-9:
Das Interview und die klassischen Gütekriterien**

Quelle: „Das verhaltensdiagnostische Interview“ von Lutz (1978, 23-24)

„In den Bereichen, in denen die klassischen Gütekriterien anwendbar sind, ... muß sich ein Interview diesen Kriterien stellen. Müssen die Testgütekriterien schlecht ausfallen, dann sind weniger die einzelnen Kennwerte von Gütekriterien von Interesse, als die Variablen, die diese Kennwerte beeinflussen... Daten aus Interviews sollen, wann immer dies möglich ist, objektiviert und validiert werden. Hierzu können Checklisten und Aufzeichnungslisten herangezogen werden..

Der Interviewverlauf sollte überprüfbar gemacht werden. Hierzu bieten sich zwei Möglichkeiten an. Zunächst muß sichergestellt sein, daß keine Fragenbereiche übersehen wurden... Durch Tonband- und Videokontrollen sollen Interviews durch Dritte (kollegiale Experten) auf innere Konsistenz geprüft werden. Kompetente Beurteiler sollen also überprüfen, ob innerhalb eines bestimmten theoretischen Konzepts nach dem bisher zur Verfügung stehenden Wissen korrekt gearbeitet wurde.

Der Therapieansatz kann als Validierung der diagnostischen Hypothese aufgefaßt werden... Durch therapiebegleitende diagnostische Maßnahmen wird ständig überprüft, ob das Ergebnis der Diagnostik, der therapeutische Ansatz, ‚valide‘ ist.“

Untersuchungen zu den Gütekriterien

Von der Frage der Modellverträglichkeit wenden wir uns Untersuchungen zu, welche die Applizierbarkeit der klassischen Testtheorie auf explorative Daten voraussetzen.

Keine Studie, die referiert wird, entspricht vollständig dem Paradigma eines ‚klassischen‘ Tests - auch wenn der Autor die partielle Inkompatibilität ‚übersieht‘.

Es liegt eine Fülle von Untersuchungen vor. Wir verweisen nur auf Publikationen, die seit Anfang der 80iger Jahre erschienen sind. *Gesichtet und zusammengestellt wurden sie von Westhofen (1991) und von Küpper (1993).*

Was leistet eine Auflistung solcher Studien?

- Sie *beweist nicht* (und kann nicht beweisen), daß ein ‚Gespräch‘ im Einzelfall objektiv, reliabel und valide ist.
- Die Beispiel verweisen jedoch auf *die prinzipielle Möglichkeit*, daß sich explorativ gewonnene Informationen - *in Grenzen* - als objektiv, reliabel und valide erweisen, sogar nach dem Kanon der klassischen Testtheorie.

Hinweis zur Auswahl: Es finden sich hohe und niedrige Werte für jedes der drei Gütekriterien. Folglich lassen sich die Studien einmal so zusammenstellen, daß *sie für* eine Meßqualität explorativer Daten, ein andermal so, daß sie *dagegen* sprechen. - Auf diese Gefahr sei nur verwiesen. Meiden könnte sie

nur eine vollständige Auflistung einschlägiger Publikationen. Diese Absicht kann sich mit den folgenden Beispielen nicht verbinden.

Unser Vorgehen: Wir zitieren für jedes Gütekriterium nur zehn Studien. Die Auswahl orientiert sich dabei an dem Ziel, möglichst unterschiedliche Ansätze zu demonstrieren.

Die Zuordnung von Gütekriterien und ‚Gespräch‘ sei in der Reihenfolge behandelt:

- explorative Daten und Objektivität (8.3.1),
- explorative Daten und Reliabilität (8.3.2),
- explorative Daten und Validität (8.3.3).

8.3.1 Explorative Daten und Objektivität

Cannell und Kahn legen ihrer ‚Philosophie‘ des Interviews folgende Vorüberlegungen zugrunde: Ein ‚Gespräch‘ lasse sich nur verstehen als eine Interaktion zwischen (mindestens) zwei Personen. Eine solche Interaktion sei immer auch dadurch charakterisiert, daß die beteiligten Partner einander beeinflussen. Darum sei es widersinnig zu erwarten, daß der Befrager bei Anamnese, Exploration oder Interview auf den Befragten keinerlei Einfluß ausübe. Ebenso sei es widersinnig zu erwarten, daß umgekehrt der Befragte in keinerlei Weise auf den Befrager einwirke. Für explorative Daten darum eine Unabhängigkeit der Erhebung zu verlangen, wie sie das Kriterium der Objektivität für Tests umschreibe, sei ‚von der Sache her‘ unangemessen (1968, 537-539).

Wenn man diese Vorüberlegung ernstnimmt: Wie weit ist es dann ‚trotzdem‘ gestattet, von einer Objektivität der Durchführung, der Auswertung und der Interpretation zu sprechen?

- Bei einer *standardisierten Anamnese oder Exploration* lassen sich *Auswertung* und *Interpretation* annähernd so regeln, wie es die klassische Testtheorie fordert. Für die *Durchführung* indessen läßt sich eine Gleichheit der ‚Bedingungen‘ so wenig garantieren wie bei einem Test.
- Bei *halb-, erst recht bei unstandardisiertem ‚Gespräch‘* ist die Sachlage komplizierter:
 - ⇒ Bei der *Durchführung* gilt: Unterschiedlicher Wortlaut der Fragen kann unterschiedliche Antworten hervorrufen. Wichtiger als das Anliegen, gleichlautende Fragen zu stellen, ist der Versuch, Fragen zu formulieren, die bedeutungsäquivalente Antworten evozieren.
 - ⇒ Bei der *Auswertung* dürfte gelten: Eine eindeutige Formulierung der Auswertekategorien und ein ausreichendes Training der Auswerter ermöglicht als Ergebnis einen hohen intersubjektiven Konsens.
 - ⇒ Für die *Interpretation* dürfte das gleiche gelten: Wohl definierte Kategorien können gleiches Verständnis herstellen.

Kasten 8-10 referiert zehn Studien - ausgewählt aus der Vielfalt von Untersuchungen, die Westhofen (1991) und Küpper (1993) gesichtet haben.

Kontrolliert wird durchgängig nur die Auswerter-Übereinstimmung. (Eine Ausnahme machen Rutter & Graham (1968, 565), die auch die Durchführungssituation zu kontrollieren versuchen - diese Studie wird hier nicht zitiert.)

Kasten 8-10:
Zur Objektivität explorativer Daten/Untersuchungen aus dem letzten Dezennium
k = Kappa, ein Übereinstimmungsmaß

Autoren/Jahr	Fragestellung	Maß
Frances et al. (1984)	Übereinstimmung zwischen zwei Auswertern	k = 0.45 bis k = 0.79
Hurt et al. (1984)	Übereinstimmung zwischen Interviewer und Patient	k = 0.89
Winslow et al. (1985)	Übereinstimmung zwischen zwei Psychiatern	k = 0.48
Erdmann et al. (1987)	Auswerter-Übereinstimmung zwischen Laien und Experten	k = 0.40 bis k = 0.80
Weyerer et al. (1988)	Übereinstimmung zwischen englischem und deutschem Interviewer	k = 0.73
Steinberg et al. (1990)	Übereinstimmung zwischen zwei Auswertern, denen ein Video vorgegeben war	k = 0.92
Leeb et al. (1991)	Auswerter-Übereinstimmung zwischen bilingual trainierten Interviewern	k = 0.78
Skre et al. (1991)	Übereinstimmung zwischen drei Auswertern	k = -0.30 bis k = 0.94
Wittchen et al. (1991)	Auswerter-Übereinstimmung zwischen Beobachter und Interviewer	k > 0.90
Williams et al. (1992)	Übereinstimmung zwischen mehreren Ratern	k = 0.68

8.3.2 Explorative Daten und Reliabilität

Wie weit lassen sich die Paradigmen der Retest-, der Paralleltest- und der Halbierungsreliabilität oder der Konsistenz auf explorative Daten anwenden?

Retest- und Paralleltestreliabilität:

- Bei völlig standardisierten Anamnesen, Explorationen, Interviews lassen sich wiederholte oder parallele Durchführungen annähernd so regeln, wie es die klassische Testtheorie fordert.
- Bei halb-, erst recht bei unstandardisierten ‚Gesprächen‘ ist die Sachlage erheblich schwieriger. Bei wiederholter oderparalleler Durchführung kann unterschiedlicher Wortlaut der Fragen auch bei denselben Personen ein unterschiedliches Verständnis wecken und unterschiedliche Antworten hervorrufen.

Halbierungsreliabilität und Konsistenz:

- *Halbierungsreliabilität:* Anamnese, Exploration, Interview in ihrer Gesamtheit lassen sich so gut wie nie in genau gleiche Hälften aufteilen.

Konsistenz: Es macht in der Regel keinen Sinn zu erwarten, daß die Fragen und Antworten eines ‚Gesprächs‘ in ihrer Gesamtheit homogen seien. Somit ist es sinnlos, für ein ‚Gespräch‘ in seiner Gesamtheit einen Wert für Konsistenz zu schätzen.

Dagegen kann es sinnvoll sein, für kleinere Abschnitte Einheiten zu bilden, von denen gilt:

- ⇒ Diese Einheiten sind in sich so homogen, daß sie die Berechnung von *Konsistenz* erlauben.
- ⇒ Diese Einheiten sind so parallel gestaltet, daß sie die Berechnung von *Halbierungsreliabilität* ermöglichen.

Kasten 8-11 bringt ein Beispiel, in dem für eine kleine Explorationseinheit Konsistenz berechnet wurde.

Kasten 8-11:
Zusammenfassung von acht Aussagen eines ‚Gesprächs‘

Quelle: „Zur Situation von Frauen in Altersheimen“, Fisseni (1974).

Das Beispiel betrifft die Auswertekategorie „Kontakte“: siehe den Kommentar unten!

Auswertung: Acht Skalenwerte geben an, wie häufig ein Verhalten auftritt	1 Wenig Kontakt	2	3	4	5	6	7	8 Reger Kontakt
(1) Gefühl des Alleinseins								
(2) Kontakte zu eigenen Kindern zu Angehörigen								
(3) Kontakte im Heim: zu Spielkreis, Plauderstunde, Skatrunde usw.								
(4) Persönlicher Kontakt zu einzelnen Bekannten im Heim								
(5) Kontakte zu Bekannten außerhalb des Heimes								
(6) Persönlicher Briefverkehr								
(7) Regelmäßige telefonische Kontakte								
(8) Allgemeine Kontaktbereitschaft im Heim								

Kommentar zu Kasten 8-9: (1) Mit 237 Frauen, die in Altersheimen wohnten, wurden Explorationen geführt und auf Tonband mitgeschnitten. - (2) Die Explorationen wurden wörtlich transkribiert. - (3) Drei Auswerter kodierten unabhängig voneinander die transkribierten Texte. Das Beispiel betrifft die Auswertekategorie „Kontakte“. - (4) Die Auswerter-Objektivität betrug $r = 0.92$. - (5) Die Konsistenz lag bei $r_k = 0.74$.

Erfahrungsregeln zur Reliabilität

Anger hat einige Faustregeln gesammelt, die Bedingungen angeben, unter denen die Reliabilität von ‚Gesprächen‘ ansteigt. (Er bezieht sich nicht auf eine bestimmte Reliabilitätsart.) - Kasten 8-12 führt die Regeln auf.

Kasten 8-12:
Erfahrungsregeln zur Reliabilität von Befragungen (Anger, 1969, 608)

Formal gilt:
- Standardisierte Befragungen sind zuverlässiger als unstandardisierte.
- Batterien von Fragen zu demselben Thema liefern zuverlässigere Informationen als Einzelfragen.
- Doppelbefragungen durch denselben Befrager erbringen zuverlässigere Ergebnisse als Befragungen durch wechselnde Befrager.
- Globalere Auskünfte sind zuverlässiger als Einzelaussagen.
Inhaltlich gilt:
- Fakten werden zuverlässiger referiert als Meinungen.
- Über die Gegenwart wird zuverlässiger berichtet als über die Vergangenheit.
- Ereignisse, die den Befragten persönlich betreffen, werden zuverlässiger berichtet als neutrale Sachverhalte.
- Qualitative Angaben sind zuverlässiger als quantitative Häufigkeiten.

Beispiele empirischer Untersuchungen: Kasten 8-13 referiert zehn Studien zur Reliabilität - ausgewählt aus der Vielfalt von Untersuchungen, die Westhofen (1991) und Küpper (1993) gesichtet haben.

Zwar überwiegt das Retest-Paradigma, doch werden auch Konsistenzschätzungen berichtet.

Kasten 8-13:
Zur Reliabilität explorativer Daten/Untersuchungen aus dem letzten Dezennium
r = Produktmoment-Korrelation. a = Alpha-Koeffizient nach Cronbach (S. 84)

Autor/Jahr	Reliabilitätsmodus	Maß
Links et al. (1985)	Retest-Reliabilität (Intervall: 3 Monate)	r = 0.78
Hurt et al. (1986)	Interne Konsistenz	r = 0.75
Hodge et al. (1988)	Retest-Reliabilität (Intervall: 3 Monate)	r = 0.91 r = 0.79
Kolko & Kazdin (1989)	Interne Konsistenz Retest-Reliabilität (Intervall: 4 Wochen)	a = 0.43 bis a = 0.85
Gross et al. (1990)	Retest-Reliabilität (Intervall: 2 Wochen)	r = 0.99
Gunderson et al. (1990)	Interne Konsistenz	a = 0.60-0.81
Nowotny et al. (1990)	Split-Half Reliabilität Interne Konsistenz	r = 0.95 r = 0.91

Weyerer et al. (1990)	Retest-Reliabilität (Intervall: 1-2 Wochen) Interne Konsistenz	$r = 0.83$ $r = 0.62$ bis $r = 0.77$
Rennen-Aillhoff (1991)	Retest-Reliabilität (Intervall: einige Tage bis wenige Wochen)	$r = 0.67$ bis $r = 0.85$
Mc Donald et al. (1991)	Split-Half Reliabilität	$r = 0.94$ $r = 0.72$

8.3.3 Explorative Daten und Validität

Welche Validitätsart läßt sich sinnvoll auf explorative Daten anwenden?

Inhaltliche Validität läßt sich für Anamnese, Exploration, Interview in ähnlicher Weise ermitteln wie für einen klassischen Test.

Allerdings wird ein Autor der eine ‚klassische‘ Inhaltsvalidierung plant, mit einem speziellen Problem konfrontiert: Explorative Informationen betreffen in der Regel auch Verhaltensprozesse. Darf der Autor solche Prozesse validieren nach einem Paradigma, welches das Invarianzkonzept der klassischen Testtheorie impliziert?

Kriterienbezogene Validität läßt sich ebenfalls anwenden, sowohl in Form von Übereinstimmungs- wie auch von Vorhersagevalidität. Soll sie in einem Koeffizienten ausgedrückt werden, so nötigt sie dazu, die Gesprächsinformationen in ein quantitatives Kategoriensystem zu übertragen und das Ergebnis zusammenzufassen (analog zu einem Test-Score).

Doch sollte ein Untersucher sich bewußt sein, daß die explorierten Inhalte ihn nötigen könnten, einen Begriff von Validität anzuwenden, der außerhalb des Invarianzkonzeptes der klassischen Testtheorie liegt.

Konstruktvalidität erfordert die Schritte, wie sie auch bei einer ‚klassischen‘ Testvalidierung nötig sind: semantische und statistische Analyse(n), vielfaltige Vergleiche innerhalb eines nomologischen Netzes.

Wieder ist plausibel: Für ein *standardisiertes ‚Gespräch‘* lassen sich die drei Validitätsarten leichter realisieren als für ein *halb- oder ein unstandardisiertes ‚Gespräch‘*.

Kommunikative Validität / Handlungsvalidität - ein Hinweis

Für explorative Daten haben sich zwei Varianten der kriteriumsorientierten Validität bewährt (Lechler, 1982, 243; Wahl, 1982, 259). Verbale Daten können ein doppeltes Ziel haben:

- Sie sollen Auskunft geben über die kognitive Repräsentanz von Probanden. Man spricht von *Kommunikativer Validität*.

- Explorative Daten sollen Auskunft geben über tatsächliches Verhalten. Man spricht von *Handlungsvalidität*.

Kommunikative Validierung besteht darin festzustellen, ob der Inhalt explorativ gewonnener Informationen übereinstimmt mit den kognitiv repräsentierten Sachverhalten des Befragten.

Diese Art der Validierung dürfte angemessen sein für Angaben über Intentionen, Interessen, Zukunftsvorstellungen, Wünsche, Gefühle, soweit die handlungsleitende Funktion außer Betracht bleibt.

Handlungsvalidierung besteht darin festzustellen, wie weit verbale Daten Zusammenhänge zu empirisch beobachtbaren Verhaltensweisen des Probanden signalisieren. Drei Hinweise zur Möglichkeit, die Validität des Zusammenhangs zu erweisen:

- Erstens können gegenwärtige Informationen *verglichen* werden mit Daten aus der *Vergangenheit*, die ihrerseits validiert sind.
- Zweitens können gegenwärtige Informationen *in einen aktuellen Kontext* plazierte werden, der seinerseits als valide gilt.
- Drittens lassen sich verbale Daten, die gegenwärtig erhoben werden, als *Prädiktoren zukünftigen Verhaltens* verwenden und so validieren.

Für alle drei Fälle handlungsbezogener Validierung liefert die Literatur eine Fülle von Beispielen. Als Kriterien dienen unterschiedliche Datenträger:

- *schulische oder forensische Akten* (Meyerhoff & Dony, 1970; Walsh, W. B., 1967, Walsh, V.R., 1976),
- *validierte Instrumente*, vor allem Fragebögen (Climent et al., 1975; Fisseni, 1974; Sines, 1959; Soskin, 1959),
- *schließlich Expertenurteile* (Gunderson & Kapfer, 1966; Landy, 1976; Walsh, V. R., 1976).

Kasten 8-14 bringt zwei Beispiele, das erste für die Kommunikative Validierung, das zweite für die Handlungsvalidierung.

Kasten 8-14:

Kommunikative Validierung und Handlungsvalidierung: Beispiele

(1) Kommunikative Validierung:

Mit 50 Medizinstudenten wurden Explorationen geführt zum Thema Selbstbild. Das ‚Gespräch‘ wurde auf Tonband mitgeschnitten. Es folgten zwei Auswertungsschritte:

- Zwei Auswerter ordneten unabhängig voneinander die Explorationsaussagen neun Selbstbildkategorien zu: *Fremdratings*.
- Jedem Probanden wurde seine Exploration vorgespielt, er stufte dann seine Aussagen auf den neun Selbstbildkategorien ein: *Selbstratings*.

Die Fremd- und Selbstratings wurden korreliert. Die Koeffizienten lagen im Durchschnitt bei $r = 0.67$; der niedrigste Wert betrug $r = 0.35$, der höchste $r = 0.75$.

Quelle: „Selbst- und Fremdauswertung von Explorationen zur Erfassung des Selbstbildes“, Kalinowsky-Czech (1984).

(2) Handlungsvalidierung:

Bei 220 Senioren, Männern und Frauen, wurden subjektive und objektive Angaben zum Gesundheitszustand erhoben.

- *Subjektive Angaben:* Mit den Probanden wurden Explorationen geführt über die Frage, wie gut sie ihren Gesundheitszustand einschätzten. Die Urteile wurden in Ratings übertragen (9 Stufen).
- *Objektive Angaben:* Alle Probanden wurden von einem Internisten untersucht. Das Gesamturteil wurde in einen Score zusammengefaßt (5 Stufen).

Zwischen beiden Urteilsreihen ergab sich ein Zusammenhang von $r = 0.31$ bei den Männern und von $r = 0.20$ bei den Frauen.

Quelle: „Psychologischer Status, subjektiver Gesundheitszustand, internistischer Befund“, Lehr, Thomae & Schmitz-Scherzer (1972).

Kasten 8-15 referiert zehn Untersuchungen zur Validität - ausgewählt aus der Recherche von Westhofen (1991) und Küpper (1993), ergänzt um Studien aus einer Arbeit von Fisseni, Olbrich, Halsig, Mailahn und Ittner (1993).

Kasten 8-15:**Zur Validität von Anamnese, Exploration, Interview/Untersuchungen aus dem letzten Dezennium**

r: Produktmoment-Korrelation, R: Multiple Korrelation,
k: Kappa, ein Übereinstimmungsmaß

Autor/Jahr	Validitätsmodus/Methoden	Maß
Halsig (1984)	Kriteriumsbezogene Validität: Vorhersage von akademischem Prüfungserfolg (Prädiktoren: Copingstrategien)	R = 0.42 bis R = 0.62
Hurt et al. (1986)	Kriteriumsbezogene Validität: Übereinstimmung zwischen Exploration und klinischer Diagnose	r = 0.63 bis r = 0.92
Vrana et al. (1986)	Kriteriumsbezogene Validität: Übereinstimmung zwischen standardisiertem Interview und Fragebogen	r = 0.89
Weidmann (1987)	Kriteriumsbezogene Validität: Vorhersage von akademischem Prüfungserfolg (Prädiktoren: Arbeitsstil und Selbstbild)	R = 0.74 bis R = 0.77
Hodge et al. (1988)	Konstruktvalidität: (A) Konvergente Validität (B) Diskriminante Validität	(A) r = 0.68 bis r = 0.77 (B) r = 0.06 bis r = 0.20
Lane et al. (1990)	Kriteriumsbezogene Validität: Übereinstimmung zwischen standardisiertem Interview und klinischer Diagnose (nach DSM-III)*	k = 0.68 bis k = 0.89
Nowotny et al. (1990)	Inhaltsvalidität	Faktorenanalyse: 5 Faktoren erklären 75 % Varianz
Weyerer et al. (1990)	Kriteriumsbezogene Validität: Übereinstimmung zwischen standardisiertem Interview und klinischer Diagnose	r = 0.69 bis r = 0.94

* DMS-III: Das revidierte Diagnostische und Statistische Manual Psychischer Störungen der American Psychiatric Association (1987).

Steinberg et al. (1990)	Diskriminante Validität: Varianzanalyse zwischen zwei Gruppen psychiatrischer Patienten und einer Gruppe Gesunder	Signifikante Differenzen auf dem 5 %-Niveau
Fisseni, Olbrich et al. (1993)	Kriteriumsbezogene Validität: Vorhersage von akademischem Prüfungserfolg (Prädiktoren Hochschullehrer-Urteile und Explorations-Urteile)	R = 0.58 bis R = 0.87

Resümee zu Teilkapitel 8.3: Studien zur Meßqualität explorativer Daten ergeben kein einheitliches Bild, die Ergebnisse divergieren. Nur mit einer Vielzahl von Differenzierungen läßt sich ein Urteil rechtfertigen. - Insgesamt läßt sich festhalten, die ‚explorative Methode‘ *ermögliche* die Gewinnung objektiver, reliabler und valider Informationen, verbürge diese drei Qualitäten aber nicht. Daraus folgt, daß die ‚Güte‘ explorativer Daten in jedem Einzelfalle zu belegen ist - wie bei jedem ‚klassischen‘ Test.

In der Einzelfalldiagnostik ließe sich eine Analogie zur Multitrait-Multimethod-Validierung bilden: Dieselben Merkmale sollten mit unterschiedlichen Verfahrensklassen erfaßt werden, so daß sich der Bedeutungshof von Merkmalen eingrenzen (Konvergenz) und von anderen Merkmalen abheben läßt (Diskriminanz).

„Als Fazit kann festgehalten werden: Auf Selbsturteile grundsätzlich zu verzichten bedeutet, Erkenntnismöglichkeiten zu vergeben. Doch dürfte es wohl weniger ihr alleiniger Einsatz als vielmehr ihre Kombination und mehr noch ihr Abgleich mit anderen Quellen und Verfahren sein, der Selbsturteile diagnostisch ergiebig macht“ (Esser, 1995, 454).

8.4 Zusammenfassung zu Kapitel 8

Befragung, Exploration, Interview oder Erhebung einer Anamnese bezeichnen eine Vorgehensweise der Informationssuche, bei der durch gezielte Fragen der Proband zu Angaben über sich und sein Umfeld angeregt werden soll. Als Oberbegriff dient die Bezeichnung „Gespräch“.

Im „Gespräch“ läuft ein vielschichtiger Prozeß ab: Er schließt Informationsverarbeitung und soziale Interaktionen ebenso ein wie lerntheoretische Prozesse der Belohnung und Bestrafung oder tiefenpsychologisch deutbare Phänomene der Übertragung und Gegenübertragung.

Das Ergebnis eines Gespräches hängt mit ab von der Form, in der es geführt wird. Im standardisierten Gespräch sind Inhalt und Reihenfolge der Fragen und der Antworten vorklassifiziert. Im unstandardisierten Gespräch bleibt es dem Befrager und Befragten überlassen, wie sie Fragen und Antworten gestalten. Häufigste Form in der Diagnostik dürfte eine Mischform sein, das sogenannte halbstandardisierte Gespräch: Ein Teil der Fragen und der Aus-

wertekategorien ist wörtlich vorgegeben; aber einen Teil der Fragen kann der Befrager im Gespräch frei formulieren und sie später in neu konzipierten Auswertekategorien einordnen.

Die Form der Fragen, die gestellt werden, bestimmt den Verlauf des Gespräches mit:

1. Geschlossene Frage geben dem Probanden Antwortoptionen vor. Offene Fragen stellen dem Probanden die Antwort völlig frei.
2. Eine direkte Frage benennt unmittelbar den Gegenstand des Gespräches. Eine indirekte Frage zielt auf ein Bedeutungsumfeld, von dem der ‚eigentliche Gegenstand‘ erreicht werden soll.

Standardisierte Gespräche liefern mit ihrer Fixierung auch ihre Auswertung. Für halb- und unstandardisierte Gespräche gibt es vielfältige Möglichkeiten, Gespräche auszuwerten. - Erstens kann das Gespräch als Original auf Video oder Tonband festgehalten werden. - Zweitens kann das Gespräch auf unterschiedliche Weise zusammengefaßt werden: Zum einen kann das Gespräch in verbale und numerische Schemata übertragen werden. Zum anderen kann es, thematisch gegliedert, zu einem fortlaufenden Text verarbeitet werden.

In der Diagnostik, vor allem im Einzelfall, ist das Gespräch unverzichtbar, wenn die Individualität und Subjektivität des Probanden für das diagnostische Problem erschlossen werden sollen.

Bei der Intervention, können Anamnese, Exploration, Interview Dienste übernehmen für alle Varianten, in denen Intervention auftritt. Im interventiven Kontext sind Gespräche jedoch nie isoliert zu betrachten, sondern immer zu sehen im Verbund mit anderen Instrumenten.

Bei der Exploration ist mit Fehlern zu rechnen: erstens mit „allgemeinen Fehlern“ wie „zentraler Tendenz“ oder „Hof-“ und „Positionseffekt“; zweitens mit speziellen Fehlern wie nonverbaler Belohnung oder Bestrafung des Probanden, ebenso mit selektiver Registrierung oder Auswertung.

Was die Gütekriterien angeht, so wird dem Problem nur eine Bewertung gerecht, die sehr viele Differenzierungen einschließt. Zunächst ist zu fragen, wie weit Gespräch und klassische Testtheorie modellverträglich sind. Sodann ist zu klären, wie weit sich ein einzelnes Gespräch - in seiner Gesamtheit oder in kleineren Einheiten - als objektiv, reliabel und valide erweisen läßt.

8.5 Kontrollfragen zu Kapitel 8

- Definition.
- Arten.
- Arten von Fragen.
- Formulierung von Fragen.

- Anwendung in der Diagnostik.
- Anwendung in der Intervention.
- Gespräch und klassische Testtheorie.

Teil III

Spezielle Einzelverfahren

In der diagnostisch-interventiven Situation muß der Psychologe nicht nur über ‚Grundkenntnisse‘ verfügen, wie Teil II sie vermitteln soll; er muß auch spezielle Einzelverfahren handhaben können.

Als ‚spezielle Einzelverfahren‘ seien - in Entsprechung zu Brickenkamp (1975, 13) - drei große Klassen genannt: Leistungs- und Persönlichkeitstests sowie Persönlichkeits-Entfaltungungsverfahren (die auch projektive Verfahren heißen).

Diese Dreiteilung ist eher eine Aufzählung als eine Gliederung unter einem einheitlichen Gesichtspunkt. Denn wie sich zeigen wird, bezeichnet ‚Leistung‘ (bei Leistungstests) den Gegenstand der Messung. -Dagegen bezeichnet ‚Persönlichkeit‘ (bei Persönlichkeitstests) keineswegs den Gegenstand der Messung; gemessen wird nicht die Gesamtperson, sondern ein Teilbereich: einige ihrer Merkmale. -Die Benennung ‚Entfaltung‘ (bei Entfaltungsverfahren) oder das Adjektiv ‚projektiv‘ (bei projektiven Verfahren) bezeichnen ein Begleitphänomen der Messung; der Anwender hofft, daß der Proband - bei Bearbeitung der Aufgaben - seine Wünsche, Bedürfnisse, Vorstellungen in den ‚Gestalten‘ entfaltet oder in sie hinein ‚projiziert‘. Was ‚Entfaltung‘ oder ‚Projektion‘ dann besagt, muß eigens geklärt werden.

Die Dreiteilung (Leistungstests, Persönlichkeitstests, projektive Verfahren) sei beibehalten, weil sie Verfahrensklassen voneinander abhebt, die für die Praxis wichtig sind.

Ein Untersucher, der diese Verfahren sachgerecht einsetzen will, muß über Grundkenntnisse verfügen, wie Teil II sie behandelt:

- Will er Leistungs- und Persönlichkeitstests verwenden, muß er die Testtheorie(n) kennen; darüber hinaus sollte er - für die Rückmeldung an den Probanden - etwas verstehen von Gesprächsführung und Verhaltensbeobachtung.
- Will er projektive Verfahren heranziehen, muß er in der Lage sein, Verhalten zu beobachten und Gespräche zu führen; will er auch die Einwände gewichten, welche Testtheoretiker erheben, muß er mit ihren Konzepten vertraut sein.

Den in einer Klasse zusammengefaßten Verfahren liegt keine einheitliche Konzeption zugrunde, weder den Leistungs- oder Persönlichkeitstests noch den projektiven Verfahren. Was die Zusammenfassung unter einen Oberbegriff begünstigt, sind eher technische Regeln der Herstellung als einheitliche Systeme ihrer theoretischen Konzeption.

Die drei Klassen seien zuerst tabellarisch miteinander verglichen, dann einzeln charakterisiert. Zunächst seien in Kasten III-1 Leistungs- und Persönlichkeitstests einander gegenübergestellt.

Kasten III-1:
Tabellarischer Vergleich von Leistungs- und Persönlichkeitstests

<i>Leistungstest</i>	<i>Persönlichkeitstest</i>
Konstruktion: nach Regeln einer <i>Testtheorie</i> . Erfassung des Personenmerkmals durch <i>Performanz</i> : Das Personenmerkmal wird realisiert (nicht nur beschrieben). Items werden bearbeitet, „soweit die <i>Fähigkeit reicht</i> “. Evoziert werden soll (demnach) <i>maximales</i> Verhalten. Den Items sind Richtig-Falsch-Lösungen zugeordnet: Zufolge einer <i>sach-immanenten Logik</i> der Aufgaben lassen sich die Antworten als ‚richtig‘ oder ‚falsch‘ klassifizieren.	Konstruktion: nach Regeln einer <i>Testtheorie</i> . Erfassung des Personenmerkmals durch <i>Abruf der kognitiven Repräsentanz</i> : Das Personenmerkmal wird beschrieben (nicht realisiert). In der Regel werden <i>alle Items</i> beantwortet. Die Beschreibung soll <i>typisches</i> Verhalten wiedergeben. Den Items werden keine Richtig-Falsch-Lösungen zugeordnet: Die Antworten ‚Stimmt‘ und ‚Stimmt nicht‘ werden <i>durch den Test-autorgepolt</i> . (Ein Item kann, mit unterschiedlicher Polung, als Indikator für verschiedene Merkmale stehen.)

Leistungs- und Persönlichkeitstests seien nun zusammengenommen und in Kasten III-2 gemeinsam den ‚projektiven Verfahren‘ gegenübergestellt.

Kasten III-2:
Tabellarischer Vergleich von projektiven Verfahren mit Persönlichkeits- und Leistungstests

<i>Leistungs- und Persönlichkeitstests</i>	<i>Projektive Verfahren</i>
<i>Reizmuster</i> (Items) werden vorgegeben als ‚geschlossene Gestalten‘. Vor der Testung wird festgelegt, welche Antworten zugelassen werden. Im Idealfall gibt es nur <i>vorklassifizierte Antworten</i> . Die Verfahren werden <i>psychometrisch vorerprobt</i> und sollten normiert sein. Die Verhaltens erfassung läßt sich charakterisieren durch die Gütekriterien der (klassischen) Testtheorie wie Objektivität, Reliabilität, Validität.	<i>Reizmuster</i> werden vorgegeben als ‚offene Gestalten‘ (bei Formdeuteverfahren als ‚Kleckse‘, bei thematischen Verfahren als ‚soziale Szenen‘). Die Antworten werden nicht vorklassifiziert, sondern im Idealfall vollständig der <i>Spontaneität des Probanden</i> überlassen. Die Verfahren werden <i>psychometrisch nicht vorerprobt</i> . Gütekriterien im Sinne der klassischen Testtheorie werden der Eigenart projektiver Verfahren nur in begrenztem Maße gerecht.

<p>Die <i>Auswertung</i> ist arbeitstechnisch ökonomisch, im Idealfall vollständig objektiv.</p> <p>Der Test-Score wird verglichen mit Kennwerten, den sogenannten <i>Normen</i>, die an einer eigenen Normstichprobe ermittelt wurden.</p> <p>Die Verfahren setzen nicht voraus, daß der Anwender und Auswerter eigens für die Datenerhebung trainiert wird, sehr wohl aber einen <i>Fachmann, der Testtheorie(n)</i> kennt.</p> <p>Zur Interpretation der Ergebnisse bedarf der Untersucher <i>keiner speziellen Erfahrung</i> - <i>sofern er die Regeln</i> der Vorgabe (Instruktion) und der Auswertung strikt <i>befolgt</i>.</p> <p>Den Verfahren entspricht eher ein <i>Strukturmodell der Persönlichkeit</i>: Die klassische Testtheorie z.B. setzt mit dem ‚wahren Wert‘ eine ‚stabile Eigenschaft‘ voraus.</p>	<p>Die <i>Auswertung</i> ist zeitaufwendig, durch <i>Training</i> jedoch in einem hohen Grade objektivierbar.</p> <p>Die Resultate werden nach sogenannten „<i>Deutehypothesen</i>“ interpretiert. Normen gibt es höchstens in dem Sinne <i>grober Richtwerte</i> - ausgenommen sehr wenige Verfahren.</p> <p>Die Verfahren setzen einen Untersucher voraus, der eingehend <i>in der Erhebungs- und Auswertungstechnik</i> trainiert worden ist (und die Theorie projektiver Verfahren kennt).</p> <p>Eine „kunstgerechte“ Interpretation der Ergebnisse setzt <i>ein hohes Maß an Erfahrung</i> bei dem Anwender voraus - er sollte ständig unter Supervision stehen.</p> <p>Den Verfahren liegt eher ein <i>Prozeßmodell der Persönlichkeit</i> zugrunde.</p>
--	---

Nach dieser tabellarischen Gegenüberstellung seien die drei Verfahrensklassen einzeln charakterisiert.

Leistungstests sind Verfahren, die nach den Regeln der (oder einer) Testtheorie konstruiert werden. In der Testsituation evozieren sie Verhaltensweisen, die zum Zielmerkmal gehören und es gleichsam operational definieren. Die Antworten lassen sich nach einer (den Aufgaben immanenten) logischen ‚Struktur‘ als richtig oder falsch klassifizieren. Der Test-Score wird verglichen mit Kennwerten, den sogenannten Normen, die an einer eigenen Normstichprobe ermittelt wurden. - Zu den Leistungstests gehören beispielsweise Intelligenz- oder Konzentrationsverfahren.

Persönlichkeitstests sind Verfahren, die ebenfalls nach den Regeln der (oder einer) Testtheorie erstellt werden. Aber die Verhaltensweisen, die das Zielmerkmal kennzeichnen, werden in der Testsituation nicht evoziert, sondern beschrieben: Persönlichkeitstests provozieren formalisierte Selbstbeschreibungen. Die Antworten ergeben sich nicht aus der logischen Sachrichtigkeit einer Aufgabenstellung, sondern durch eine Zuordnung des Testautors. (Der Testautor sieht Gründe, bestimmte Antworten als Indikatoren des Zielmerkmals zu interpretieren.) Wie bei Leistungstests wird der Test-Score verglichen mit Kennwerten, die an einer Normstichprobe ermittelt wurden. - Andere Bezeichnungen lauten Persönlichkeitsinventar oder (Persönlichkeits-)Fragebogen.

Schwieriger ist **es, projektive Verfahren** zu charakterisieren. Es sind Instrumente, bei denen angenommen wird, daß ihre Reize als „offene Gestalten“ - etwa als Kleckse oder soziale Szenen - Antworten hervorrufen, die den ‚Projektionen‘ und ‚Identifikationen‘ des Probanden entspringen. Aus den Projekt-

tionen und Identifikationen wird auf zugrundeliegende Persönlichkeitsmerkmale geschlossen.

Wir stellen die drei Verfahrensklassen ausführlich vor:

- zuerst die Leistungstests (Kap. 9),
- sodann die Persönlichkeitstests (Kap. 10).
- schließlich die projektiven Verfahren (Kap. 11).

9. Kapitel

Leistungstests

Kapitel 9 behandelt eine erste Klasse von Verfahren, zu deren Verständnis das Basiswissen aus Teil II gehört, vorrangig Kenntnisse der Testtheorie(n), aber auch - für die Rückmeldung an den Probanden - Vertrautheit mit Gesprächsführung und Verhaltensbeobachtung.

Was trägt ein Leistungstest zu Diagnostik und Intervention bei? In drei Abschnitten versuchen wir eine Antwort zu geben, wir behandeln:

- Allgemeine Charakteristika von Leistungstests (9.1),
- Analyse und Vergleich von Testdaten (9.2),
- Kriterien und Beispiel einer Testbewertung (9.3).

9.1 Allgemeine Charakteristika von Leistungstests

Die allgemeinen Charakteristika von Leistungstests besprechen wir in fünf Abschnitten:

- Definitionen und Abgrenzungen (9.1.1),
- Klassifikation von Leistungstests (9.1.2),
- Beitrag zu Diagnostik und Intervention (9.1.3),
- Aufgabenfeldern für Leistungstests (9.1.4),
- Resümee (9.1.5).

9.1.1 Definitionen und Abgrenzungen

Der Begriff ‚Leistungstest‘ soll Verfahren bezeichnen, die nach den Regeln der (oder einer) Testtheorie konstruiert werden und eine Stichprobe jener Verhaltensweisen erheben, die zum Zielmerkmal gehören und es gleichsam operational definieren. Die Antworten lassen sich als ‚richtig‘ oder ‚falsch‘ klassifizieren, nach einer den Aufgaben immanenten logischen Struktur. - Kasten 9-1 veranschaulicht diese Aspekte an zwei Beispielen.

Kasten 9-1:**Items aus einem Leistungstest**

*Quelle: Hamburg- Wechsler-Intelligenztest für Erwachsene
(Revision 1991, HAWIE-R, Tewes)*

Lösung hinter dem Item in Klammern.

UNTERTEST ‚Allgemeines Wissen‘ (AW):

Item 3: „In welche Himmelsrichtung fährt man, wenn man von Hamburg nach München reist?“ (Nach Süden).

Item 19: „Was ist eine Ode?“ (Feierliches Gedicht).

UNTERTEST ‚Rechnerisches Denken‘ (RD):

Item 12: „Eine Familie kauft einen gebrauchten Schrank für zwei Drittel des Neupreises. Sie bezahlt 400 DM für den Schrank. Wie hoch war der Neupreis?“ (600 DM).

Item 14: „Mit 8 Maschinen kann man eine Arbeit in 6 Tagen erledigen. Wieviel Maschinen sind nötig, um die gleiche Arbeit in einem halben Tag zu erledigen?“ (96).

Kommentar: Der Proband muß im Untertest AW Wissen abrufen, im Untertest RD (sowohl Wissen abrufen als auch) eine Dreisatzaufgabe lösen.

Normierung und Klassifikation

Kasten 9-2:**Normierung und Klassifikation**

Quelle: Lern- und Gedächtnistest, Stufe 3, Form A (LGT 3: Bäumler, 1974)

Für die Gedächtnisleistungen, die eine 18jährige Abiturientin in den 6 Untertests des LGT-3 erbracht hat, erhält sie als Test-Score eine Summe von Rohpunkten. - Anhand der Normtabellen werden die Rohpunkte in T-Werte umgewandelt (Mittelwert: 50; Standardabweichung: 10).

Die Normen erlauben unterschiedliche **Klassifikationen**, hier seien drei Unterscheidungen getroffen:

- **Durchschnittlich:** Mittelwert \pm 1 Standardabweichung,
hier also T-Wert 40-60,
- **Unterdurchschnittlich:** Leistung unter T-Wert 40,
- **Überdurchschnittlich:** Leistung über T-Wert 60.

Untertests, Rohpunkte (RP), T-Werte (T-W) und Klassifikationen seien in einer Übersicht einander zugeordnet.

<i>Untertest</i>	<i>RP</i>	<i>T-W</i>	Klassifikation
1. Stadtplan	14	42	Durchschnittlich
2. Türkisch	11	51	Durchschnittlich
3. Gegenstände	14	66	Überdurchschnittlich
4. Telefon	9	62	Überdurchschnittlich
5. Bau	7	42	Durchschnittlich
6. Zeichen	14	59	Durchschnittlich

Kommentar: Es gibt gute Gründe, die Klassifikationsgrenzen anders zu ziehen als hier in dem Beispiel. So könnte ein Untersucher als *durchschnittlich* den Bereich „Mittelwert \pm 0.5 Standardabweichung“ festlegen; das Intervall umfaßte in diesen Kasten dann die T-Werte von 45 bis 55. (Siehe Normierung, S. 118)

Der Test-Score wird mit Normen verglichen. Kasten 9-2 gibt ein Beispiel. Der Test-Score ist die Summe der Punkte für richtige Antworten. Normen sind Kennwerte, die an einer Eichstichprobe ermittelt wurden (siehe S. 111).

Begriffliche Unterscheidungen

Leistungstests sollen jene Verhaltensanteile erfassen, die sich als ‚Leistungen‘ einstufen lassen.

Um diese Leistungsanteile genauer zu charakterisieren, treffen wir einige begriffliche Unterscheidungen. Wir heben Begriffe voneinander ab, die Ähnliches umschreiben. Gemeint sind ‚Leistung‘, ‚Fähigkeit‘, ‚Fertigkeit‘, ‚Eignung‘ und ‚Begabung‘.

Welchem Zweck dient die Abgrenzung? Sie soll dafür sensibilisieren, daß in dem Konzept der Leistung unterschiedliche Aspekte enthalten sind.

In einem allgemeinen Sinn bezeichnet **Leistung** eine in einer Zeit(einheit) erbrachte Arbeit. „Vom Subjekt her gesehen muß also eine bestimmte eigene Energie aufgewandt werden, es muß ein bestimmter Kräfteinsatz möglichst zweckmäßig und dem ‚Problem der Sache‘ entsprechend erfolgen“ (Thomae, 1968, 369).

So kann Leistung den ‚Kräfteinsatz‘ bei einer erfolgreichen Bearbeitung von Items bezeichnen. Beispiele solcher ‚Leistungen‘ sind die Lösung von Aufgaben eines Tests oder die Beantwortung von Fragen eines Persönlichkeitsinventars. - In dieser Bedeutung eignet sich der Begriff nicht dazu, zwischen Persönlichkeits- und Leistungstest zu unterscheiden.

Hier sei das Konstrukt enger gefaßt. Leistung soll jene ‚Anstrengung‘ charakterisieren, die ein Proband aufbringt, um das Zielmerkmal eines Tests zu ‚realisieren‘. - Von ‚Leistungen‘ in diesem engeren Sinne kann man, mit gewissen Vorbehalten, sagen, daß sie beobachtbar sind. Nicht beobachtbar ist, was das nächste Konstrukt umschreibt.

Eine **Fähigkeit** soll die *psychischen und somatischen Bedingungen* angeben, die eine Leistung (im engeren Sinne) ermöglichen. Es sind in „in der Lebensgeschichte entstandene komplexe Eigenschaften, die . . . den Tätigkeitsvollzug steuern“ (Dorsch, 1994, 229). Fähigkeiten kann man nicht beobachten, man muß sie erschließen, Beispiele sind: Intelligenz, Konzentration, Handgeschick.

Eine **Fertigkeit** soll die zu einer Leistung notwendigen Techniken, Erfahrungen, Kenntnisse umschreiben, die durch *Übung* erworben sind.

Das gleiche wie Fähigkeit oder Fertigkeit kann auch der Begriff **Eignung** bezeichnen. Doch kann Eignung auch einen umfassenderen Sinn annehmen: die Einbettung von Fähigkeiten und Fertigkeiten in die Gesamtheit einer Person und ihres Umfeldes; dann schließt Eignung auch Motive, Einstellungen

oder Interessen ein, also etwa die Gesamtheit der ‚Erfordernisse‘ für den Beruf eines Sonderschullehrers.

Als Synonym zu Eignung wird **oft Begabung** verwandt. Das Konstrukt läßt sich aber auch enger verstehen: als *Optimum an Eignung*. Bei dieser Sprachregelung bezeichnet ‚Eignung‘ ein ‚fundamentales Können‘, ‚Begabung‘ dagegen ein ‚vollendetes Können‘.

Beispiel: *Es sei angenommen, jemand sei „musikalisch“, er habe ein Gehör das Tonhöhen, Klangfarben, Klangintensitäten klar unterscheidet, er habe ein Gespür für Rhythmik, er habe Interesse an Musikgeschichte usw. Damit ist er für ein Musikstudium **geeignet**, aber noch nicht in dem genannten Sinne für Musik auch schon begabt.*

Begabt wäre er, wenn er ein **ausgezeichnetes** Gehör hätte, ein **auffälliges** Gespür für Rhythmik, vielleicht auch **überaus** geschickte Finger und ein „**unstillbares**“ Interesse für Fragen der Musik. *Begabung soll anzeigen, daß jemand über besonders günstige Voraussetzungen für bestimmte Leistungen verfügt.*

Was ist nach den genannten Abgrenzungen Gegenstand der Leistungstests?

Leistungstests bezeichnen eine Gruppe von Verfahren, die einem Probanden eine **Leistung** abverlangen, in der Anteile des Zielmerkmals ‚realisiert‘ werden.

Aus der Leistung wird auf zugrundeliegende **Fähigkeiten** und **Fertigkeiten** geschlossen, manchmal auf eine **Eignung**, gegebenenfalls sogar auf eine besondere Eignung, auf eine **Begabung**.

Welche dieser Funktionseinheiten sind in Leistungstests erfaßt worden?

Funktionseinheiten als Gegenstand von Leistungstests

Aus unterschiedlichen Interessen wurden Verfahren entwickelt, um spezielle Funktionseinheiten zu erfassen, also *Fähigkeiten, Fertigkeiten, Eignungen, Begabungen* wie die folgenden (Sauermann, 1979, 80-81):

1. Allgemeine Leistungsfunktionen.

z. B. Konzentration, Aufmerksamkeit, Ausdauer, Arbeitstempo, Belastbarkeit, Genauigkeit;

2. Motorische und sensumotorische Fähigkeiten:

z. B. Psychomotorische Reaktionen, Auge-Hand-Koordination, Motorisches Tempo, Treffsicherheit motorischer Bewegungen;

3. Wahrnehmungsfähigkeit:

z. B. Sehfähigkeit, Farbtüchtigkeit, Hörfähigkeit, Formauffassung;

4. Gedächtnis- und Lernfähigkeit:

z. B. numerisches, verbales, visuelles Gedächtnis, kurz-, mittel-, langfristiges Gedächtnis;

5. Intellektuelle Fähigkeiten:

z. B. induktives Denken, deduktives Denken, Raumvorstellung, Wortflüssigkeit.

Jeder Funktionseinheit läßt sich eine Reihe von Leistungstests zuordnen. Welche von ihnen lassen sich zu besonderen Klassen zusammenfassen?

9.1.2 Klassifikation von Leistungstests

In den Funktionseinheiten, welche durch Leistungstests erfaßt werden, lassen sich zwei Anteile unterscheiden, allgemeine und spezielle. Nach diesen beiden Anteilen kann man Leistungstests auch einteilen.

HINWEIS: Die Einteilung in allgemeine und spezielle Leistungstests ist nur eine unter vielen denkbaren Klassifikationen. So schlagen Hiltmann (1966) oder Brickenkamp (1975, 8) andere Gliederungen vor; auf diese anderen Möglichkeiten sei nur verwiesen.

Allgemeine Leistungstests

Es gibt Verhaltensanteile, die in jeder Leistung enthalten sind (Bartenwerfer, 1983, 482). Auf diese Anteile beziehen sich allgemeine Leistungstests.

„Die allgemeinen Leistungstests erfassen hauptsächlich die Fähigkeit des Individuums, eine der richtigen Aufgabenlösung dienende angemessene ‚innere Grundlage‘ zu schaffen und über die erforderliche Zeit hinweg aufrecht zu erhalten“ (Bartenwerfer 1983, 485).

Unter den allgemeinen Leistungstests werden, aus Gründen psychologiegeschichtlicher Entwicklung, Intelligenztests als eigene Klasse gefaßt und der Klasse der ‚anderen allgemeinen Leistungstests‘ gegenübergestellt. In dieser Einteilung spiegelt sich die Rolle wider, welche ‚Intelligenz‘ in der Gesellschaft spielt, in der wir leben.

Es folgen einige Beispiele:

Intelligenztests: Die Konstruktion von Intelligenztests orientiert sich an unterschiedlichen Theorien. Drei seien erwähnt (Conrad, 1983, 108-122; Groffmann, 1982).

1. Spearman (1927) konzipierte ein „Zweifaktorenmodell“: Intelligenz gliedert sich in zwei Komponenten auf:
 - in einen allgemeinen Faktor, der in jeder Intelligenzleistung wirksam wird (der sogenannte „general factor“, Kürzel „g“), und
 - in mehrere spezielle Faktoren, die bei einzelnen kognitiven Anforderungen zusammen mit „g“ wirksam werden (sogenannte „special fac-

tors“, Kürzel „s“); spezielle Anforderungen könnten Aufgaben betreffen, die verbale oder numerische Probleme aufwerfen.

Beispiele:

- a) Der „*Figure Reasoning Test*“ (FRT): Der Test verwendet sprachfreie Items, in deren Aufbau Gesetzmäßigkeiten zu erkennen sind. Erfasst werden soll die Allgemeine Intelligenz, also der „g-Faktor“ (Daniels, 1971).
- b) Die „*Standard Progressive Matrices*“ (SPM): Der Test verwendet drei Serien A, B, C sprachfreier Aufgaben; er dient zur Erfassung der Fähigkeit zu logischem Denken, zentriert um den Generalfaktor „g“. Ergänzt werden die SPM oft durch einen Sprachtest (Raven, 1971: Deutsche Bearbeitung von Kratzmeier, 1978).

2. Thurstone (1938) konzipierte ein „*Gruppenfaktorenmodell*“: Intelligenz gliedert sich in mehrere Faktoren, die voneinander unabhängig sind. Solche sogenannten *Primärfaktoren* sind:

- v: verbal comprehension (passiver Wortschatz),
- w: verbal fluency (aktiver Wortschatz),
- n: number (rechnerisches Denken),
- s: space (räumliches Vorstellen),
- p: perceptual speed (Wahrnehmungstempo),
- r/i: reasoning oder induction (deduktives oder induktives Denken),
- m: memory (Merkfähigkeit).

Wird das Gruppenfaktorenmodell so strikt interpretiert, wie es konzipiert war, dann ist in ihm kein Platz für eine Aussage zur Gesamtintelligenz.

Beispiel: Ein Test, der sich an diesem Modell orientiert, ist das „*Leistungsprüfsystem*“ (LPS: Horn, 1983). Das Verfahren besteht aus vierzehn Untertests, von denen wenigstens je zwei die Primärfaktoren erfassen sollen.

3. Jäger, A. O. (1967) hat das Gruppenfaktorenmodell von Thurstone übernommen und zugleich modifiziert: Die *Primärfaktoren* der Intelligenz werden *hierarchisch geordnet*. Sie heißen:

- Anschauungsgebundenes Denken,
- Einfallsreichtum und Produktivität,
- Konzentrationskraft und Tempomotivation,
- Verarbeitungskapazität: Formallogisches Denken und Urteilsfähigkeit,
- Zahlengebundenes Denken,
- Sprachgebundenes Denken.

„Messungen der Allgemeinen Intelligenz (g) und Messungen der Fähigkeitsstruktur sind keine einander ausschließenden Alternativen. Allgemein gewinnt die Auffassung an Boden, daß sich ein hierarchisches Strukturmodell der Fähigkeiten, orientiert an deren Generalitätsgraden, mit g (der Allgemeinen Intelligenz) an der Spitze als zweckmäßig erweist“ (Jäger, A. O. & Althoff, 1984, 5; Althoff, 1984).

Beispiel: Der „*Wilde-Intelligenz-Test*“ (WIT) soll dieses hierarchische Strukturmodell abbilden. Fünfzehn Untertests lassen sich zu Primärfaktoren zusammenfassen und ermöglichen darüber hinaus eine Schätzung des

generellen Faktor „g“. - „Im WIT ist dementsprechend neben der Erfassung von bedeutsamen Primärfaktoren eine Zusammenfassung aller Subtestergebnisse zu einer Schätzung der Allgemeinen Intelligenz vorgesehen“ (Jäger A. O. & Althoff, 1984, 5).

Konzentrationstests: Konzentration, also eine Ausrichtung der Aufmerksamkeit auf eng umgrenzte Sachverhalte (für eine begrenzte Zeit), wird ‚auf dem Umweg‘ über drei „Operationen“ erfaßt - Durchstreichen, Rechnen, Sortieren:

1. **Durchstreichen:** Der „d2 Aufmerksamkeits-Belastungs-Test“ gibt dem Probanden vierzehn Buchstaben-Reihen vor. Für jede Reihe werden ihm zwanzig Sekunden Arbeitszeit gewährt. Seine Aufgabe ist es, jedes „d“ **durchzustreichen**, das mit zwei Strichen versehen ist. Gemessen werden Arbeitstempo und konzentrierte Sorgfalt (Brickenkamp, 1994).
2. **Rechenaufgaben lösen:** Der „Konzentrations-Leistungs-Test“ (KLT) gibt dem Probanden **Additions- und Subtraktionsaufgaben** vor. Jede Aufgabe besteht aus zwei Zeilen. Zunächst ist das Ergebnis (E) der oberen Zeile auszurechnen (E_{oben}), dann das der unteren Zeile (E_{unten}). Nun ist zu entscheiden: Ist E_{oben} kleiner als E_{unten} , dann sind beide Ergebnisse zu addieren. Sonst ist E_{unten} von E_{oben} zu subtrahieren. - Als Testzeit werden dreißig Minuten festgelegt. Ermittelt werden Leistungsmenge und Konzentrationsgute (Düker & Lienert, 1965).
3. **Sortieren:** Der „Konzentrations-Verlaufs-Test“ (KVT) gibt dem Probanden 60 Karten vor, die er nach einem bestimmten Prinzip in vier Gruppen **sortieren** soll. Gemessen werden die Bearbeitungszeit und die Fehlerzahl; eine Kombination aus Fehler und Zeit wird als Sorgfaltsleistung interpretiert (Abels, 1965).

Gedächtnistests: In mehreren Intelligenztest ist ein Untertest vorgesehen, der das kurz- und mittelfristige Gedächtnis erfassen soll. Solche Untertests enthält beispielsweise der eben genannte WIT von Jäger, A. O. und Althoff (1984) sowie der „Intelligenzstruktur-Test“ von Amthauer (1974). - Das folgende Beispiel nennt ein spezielles Verfahren:

Beispiel: Der „Lern- und Gedächtnistest, Stufe 3“ (LGT 3) mißt die Merkfähigkeit in sechs Untertests. Zu erlernen sind verbale, numerische und figurale Aufgaben. Der Proband muß sie sich einprägen und später wieder reproduzieren (Bäumler, 1974).

Spezielle Leistungstests

Es gibt auch Verhaltensanteile, die nicht in jeder Leistung enthalten sind, sondern nur bei speziellen Anforderungen abgerufen werden. Beispiele sind Aufgaben, zu deren Lösung Finger- oder Handgeschick, mechanisch-technisches Verständnis oder organisatorische Fähigkeiten benötigt werden. Tests, die diese Funktionseinheiten erfassen, heißen spezielle Leistungstests.

Es folgen einige Beispiele:

Entwicklungstests: Entwicklung kennzeichnet den menschlichen Organismus von der Vereinigung der Keimzellen bis zu ihrem Absterben. Entsprechend vielfältig und zahlreich sind die Verfahren, die dazu bestimmt sind, Entwicklung zu erfassen. - Erwähnt seien zwei „Klassiker“:

1. Der „Bühler-Hetzer-Kleinkindertest“ sieht vor, die menschliche Entwicklung vom ersten Lebensmonat bis zum siebenten Lebensjahr zu erfassen. Der Entwicklungsstand wird nach sechs Verhaltensbereichen geordnet: (1) Sinnliche Rezeption (Ansprechbarkeit der Sinnesorgane), (2) Körperbeherrschung (Steuerung der Bewegungen), (3) Soziales Verhalten (verbale, nonverbale Kontakte), (4) Materialbeherrschung (Greifen nach „Material“ in der Umwelt), (5) Geistige Produktion (Erfassen von Sinnzusammenhängen), (6) Lernen (Zeränderung des Verhaltens durch Erfahrung und Nachahmung). - (BHKT: Bühler & Hetzer, 1972).
2. Die „Lincoln-Oseretzky-Scale, Kurzform“, dient zur Prüfung der motorischen Entwicklung behinderter lernbehinderter und normaler Kinder im Alter von 5 bis 13 Jahren (LOS KF 18, Eggert, 1994).

Schultests: Pädagogische Tests, die in Schulen eingesetzt werden, können unterschiedlichen Zielen dienen. Drei dieser Ziele seien erwähnt:

- Schuleintrittstests prüfen die ‚Reife‘ für eine bestimmte Schule;
- Übertrittstests prüfen die Eignung für bestimmte Schularten;
- Schulleistungstests prüfen den Leistungsstand auf bestimmten Klassenstufen.

Tests zur Prüfung spezieller Funktionen: Viele Verfahren wurden entworfen, um so spezielle Funktionen wie Musikalität, Finger- und Handgeschick oder auch Koordination von Sensorik und Motorik zu messen.

9.1.3 Beitrag zu Diagnostik und Intervention

Leistungstests „haben sich im Netzwerk empirischer Beziehungen bewährt“ (Klauer, 1978, 7). Ihr Beitrag für Diagnostik und Intervention ergibt sich aus ihrer Eigenart, er wird den Verfahren nicht von außen zugeordnet.

Was leisten Tests für die Diagnostik?

1. Leistungstests ermöglichen es dem Diagnostiker, **ein Merkmal unter kontrollierten Bedingungen zu erfassen**. Vor allem zwei Momente können diese Kontrolle gewährleisten: erstens die Konstruktionsregeln, zweitens die Anwendungsvorschriften.
Über das Ausmaß, in dem die **Konstruktionsregeln** befolgt wurden, gibt das Testmanual Auskunft, wenn es Item- und Testkennwerte referiert (oder

nicht referiert). Diese Angaben ermöglichen einen Umfang an Kontrolle, wie ihn ‚offene‘ Verfahren (in der Regel etwa Verhaltensbeobachtung oder Exploration) nicht gewähren.

Für die **Anwendung** schreibt die Instruktion die Prozeduren vor. Von Kontrolle kann hier allerdings nur soweit die Rede sein, wie sich der Anwender an diese Vorschriften hält.

2. Leistungstests ermöglichen es dem Diagnostiker, ein **Merkmal auf der Verhaltensebene zu erfassen**. Sie referieren nicht nur über Ideen, welche die Merkmale repräsentieren sollen, sondern legen konkrete Situationen und Operationen fest, in denen das Merkmal realisiert werden kann.
3. Leistungstests ermöglichen eine **Vielzahl von Vergleichen** - Vergleiche von Individuen oder Gruppen mit Individuen oder mit Gruppen. Bezogen wird auf unterschiedliche **Klassifikationen**, welche ein Test ermöglicht, vor allem dank seiner Normierung. Es werden - um es so zu formulieren - beispielsweise soziale, epochale, regionale Koordinatensysteme angeboten, in denen sich die relative Position eines einzelnen oder einer Gruppe bestimmen läßt.

Beispiel: Der IST 70 erlaubt es, einen Probanden zu vergleichen mit Gleichaltrigen, mit Angehörigen verschiedener Bildungsgruppen oder Vertretern verschiedener Berufe (Amthauer, 1973).

4. Leistungstests ermöglichen es dem Diagnostiker im Idealfall, ein **Merkmal in einem theoretischen Kontext** zu beschreiben. Das Zielmerkmal wird in den Kontext einer Theorie gebettet und erklärt.

Beispiel: Das „Leistungsprüfsystem (LPS)“ ist konzipiert nach der Intelligenztheorie von Thurstone (Horn, 1983; siehe S. 268).

Daß jedoch speziell in der Frage eines Theoriebezugs viele Mängel vorherrschen, sei unbestritten.

5. Leistungstests können zur **Sprachregelung** innerhalb der Psychologie beitragen. Sie geben bekannt, wie Konstrukte mit zugehörigen Operationen verknüpft werden. Damit bieten sie - und das geht über Sprachregelung sogar hinaus - die Möglichkeit, die Performanz von Verhalten in den vom Test ‚geregelten‘ Situationen zu beobachten, gegebenenfalls zu korrigieren und zu verbessern.

„Intelligenztests haben nicht nur eine deskriptiv-messende Funktion, sie wirken auch präskriptiv in dem Sinne, daß sie operationale Definitionen von Intelligenz sind, also festschreiben, was intelligentes Verhalten ist“ (Fay, 1993, 278; vgl. S. 291 in diesem Buch).

Was leisten Tests für die Intervention?

Von ihrer Eigenart her können Leistungstests unterschiedliche Aufgaben übernehmen.

1. Leistungstests können helfen, **Interventionsbedarf festzustellen**. Solche Feststellungen lassen sich auch ohne Tests treffen, aber Tests bieten be-

stimmte Vorzüge: Sie umreißen ein wohldefiniertes Merkmal, sie geben - aufgrund der Normierung - empirisch abgesicherte Klassifikationen vor. Zur genaueren Bestimmung des Interventionsbedarfs gehören in der Regel weitere Angaben, etwa aus Akten, Anamnesen oder Gesprächen mit Betroffenen und ihren Bezugspersonen.

Beispiele:

- Bei sensumotorischen Funktionen können Störungen identifiziert werden, so etwa, daß bei einem Patienten nach einem Motorradunfall die Auge-Hand-Koordination verlangsamt ist.
- Bei kognitiven Fertigkeiten können Leistungstests Defizite anzeigen, so etwa, daß ein 13jähriger Schüler die dezimale Grundstruktur des Zahlensystems nicht erfaßt.
- Bei der Wahrnehmungsfähigkeit können Leistungstests Ausfälle anzeigen, so etwa, daß bei dem Klienten eines Rehabilitationszentrums das visuelle Feld erheblich eingeschränkt ist.

2. Tests oder testähnliche Verfahren können der **Prävention** oder dem **Training** dienen. Sie lassen sich in eine Sequenz interventiver Maßnahmen (Treatments) einfügen - besonders einfach dann, wenn es sich um edv-gestützte Instrumente handelt.

Beispiele:

- Bei dem Unfall-Patienten kann die Auge-Hand-Koordination an Reaktionsgeräten trainiert werden: ein solches Training wird eingefügt in andere ‚Verordnungen‘, etwa Krankengymnastik unterschiedlicher Art.
- Dem 13jährigen Schüler können Aufgaben eines Zahlentests dabei helfen, die Grundstruktur des Zahlensystems zu durchschauen und ihre numerischen Relationen zu behandeln, im Verbund mit anderen Maßnahmen.
- Bei dem Klienten des Rehabilitationszentrums kann das Sehfeld möglicherweise wieder ergänzt werden durch Training an Sehtests oder analogen Verfahren, auch hier im Verbund mit anderen Prozeduren.

3. Leistungstests oder testähnliche Verfahren können die **Bilanzierung** einer Intervention erleichtern: indem sie den Fortschritt eines Trainings konstatieren, indem sie also als Instrumente von Prozeß- oder Erfolgsmessung dienen. Erleichtert wird dieser Dienst wiederum, wenn edv-gestützte Instrumente zur Verfügung stehen.

Beispiele:

- Ob ein Auge-Hand-Training Erfolg hatte, läßt sich dokumentieren durch Vergleich der Reaktionen zu Beginn und nach bestimmten Abschnitten der Übung.
- Ebenso kann das Ergebnis des Trainings mit mathematischen Tests leicht konstatiert werden.
- Daß bei dem Klienten in der Rehabilitation das Sehfeld tatsächlich wieder ausgeweitet wurde: auch dieses Ergebnis kann ein Sehtest ‚sichtbar‘ machen.

GEGENFRAGE: Am Ende dieses Abschnittes sei auch die Gegenfrage gestreift: Was darf ein Untersucher von Leistungstests nicht erwarten? Kasten 9-3 umreißt eine Antwort.

Kasten 9-3:
Wozu sind Leistungstests nicht konzipiert?

1. Leistungstests machen nicht die ‚inneren Prozesse‘ sichtbar, denen das realisierte Verhalten entspringt. In diesem Sinne liefern sie nur eine Ergebnisdiskription, aber keine Prozeßinformation.
Für bestimmte Fragestellungen sind Informationen über einen Leistungsprozeß aufschlußreich, beispielsweise wenn ein Berufswunsch beurteilt oder ein Ehekonflikt verfolgt werden soll. Prozeßinformation liefern andere Instrumente, etwa Beobachtung, Exploration, projektive Verfahren.
2. Leistungstests beschreiben Verhalten mittels vorklassifizierter Kategorien (der Items und ihrer Zusammenfassung in ‚Merkmalen‘). Die Testperson wird in einem vorher konstruierten Koordinatensystem beurteilt. Spontanes Verhalten bleibt unberücksichtigt.
In diesem Sinne weist Testdiagnostik über sich hinaus: auf eine Erfassung von Verhaltensanteilen, die in der diagnostischen Situation zwar sichtbar werden, aber in den Testkategorien nicht vorgesehen sind. Solche Verhaltensanteile lassen sich möglicherweise in einer Verhaltensbeobachtung festhalten.
3. Vom Verhalten in der Testsituation kann nicht problemlos auf das Verhalten außerhalb der Testsituation geschlossen werden.
4. Testnormen erlauben es nur dann, die relative ‚Position‘ eines Probanden zu bestimmen, wenn der Proband den Charakteristika der Normstichprobe entspricht. Genau diese ‚Entsprechung‘ läßt sich nicht immer mit völliger Klarheit ausmachen. Darüber hinaus ist zu beachten, daß Normen vergleichsweise schnell veralten.

9.1.4 Aufgabenfelder für Leistungstests

Es seien Aufgabenfelder genannt, auf denen Leistungstests vorzügliche Beiträge erbracht haben:

In der **Berufsberatung** stellen Probanden unterschiedliche Fragen: Wozu bin ich ‚geeignet‘? Wo liegen meine ‚Fähigkeiten‘? Oder umgekehrt: Besitze ich die ‚Fähigkeiten‘ für einen bestimmten Beruf? - Zur Beantwortung solcher Fragen können Leistungstests eine Reihe nützlicher Informationen liefern.

Einen weiten Bereich, in dem eignungsdiagnostische Fragen zu klären sind, eröffnen **Industrie, Verwaltung, Bundeswehr**. Bei Aufgaben wie Personalauslese und Personalentwicklung lassen sich Eignungsfragen anhand von Test-Scores mit-entscheiden.

Ein eigenes Aufgabenfeld gibt sodann die **Verkehrspsychologie** vor, die Leistungstests heranzieht, um Probleme zu behandeln wie die Frage,

- ob speziellen Gruppen die „Zulassung zur motorisierten Straßenverkehrsteilnahme“ (der „Führerschein“) erteilt werden soll, etwa Älteren, Behinderten,

- oder ob die Fahrerlaubnis (der „Führerschein“) wiedererteilt werden soll, beispielsweise nach Entzug wegen Trunkenheit am Steuer.

Ein umfassendes Feld für eignungsdiagnostische Untersuchungen bietet auch die **Schulpsychologie**, allgemein die **Pädagogische Psychologie**. Es stellen sich Aufgaben wie Prüfung von Schul- oder Hochschulreife oder Ermittlung des Kenntnisstandes auf einzelnen Klassenstufen.

Leistungsfragen stellen sich auch im **Rehabilitationswesen**, beispielsweise wenn zu klären ist, welche ‚psychischen Funktionen‘ ausgefallen sind und ob ihr Wiedererwerb eingeübt werden kann.

HINWEIS: Die einzelnen Aufgabenfelder haben sich in einem hohen Grade verselbständigt und spezialisiert. Wer das eine Feld kennt, überblickt damit noch nicht die Anforderungen eines anderen Feldes.

9.1.5 Resümee zu Kapitel 9.1

Leistungstests erfassen Verhaltensanteile, in denen das Zielmerkmal ‚realisiert‘ werden soll. Dieser Konzeption ist affin ein persönlichkeits-theoretisches Modell, das stabile Merkmale (traits) voraussetzt - ein für viele diagnostische Fragestellungen sinnvoller Denkansatz, vor allem in Zusammenhang mit ‚eignungsdiagnostischen‘ Aufgaben.

Ein Test kann auf Verhaltensanteile zielen, die in jeder Leistung enthalten sind, dies ist Gegenstand der ‚allgemeinen Leistungstests‘. Ein Test kann aber auch Verhaltensanteile erfassen, die nur in besonderen Leistungen enthalten sind, dies ist Gegenstand der ‚speziellen Leistungstests‘.

Für die Diagnostik erbringen Leistungstests Beiträge wie (1) kontrollierte Merkmalserfassung, (2) Merkmalsermittlung auf Verhaltensebene, (3) Ermöglichung vieler Vergleiche, (4) Vorgabe eines Merkmals in einem theoretischen Netz, (5) Sprachregelung innerhalb der Psychologie.

Für die Intervention kann ein Test dazu dienen (1) ‚Orte‘ festzustellen, an denen ein Eingriff vonnöten ist, (2) Instrumente für Prävention oder Training bereitzustellen, (3) eine Bilanz für Erfolg oder Mißerfolg zu erleichtern.

Es wurden einige Aufgabenfelder genannt, auf denen sich Leistungstests bewährt haben, beispielsweise die Berufsberatung, die Verkehrspsychologie, das Rehabilitationswesen.

ERGÄNZUNG: Fähigkeiten und Fertigkeiten, Eignungen und Begabungen lassen sich nicht nur mit Leistungstests erfassen.

Hinweise geben auch andere Verfahren, beispielsweise

- die Analyse biographischer Daten, etwa amtlicher Akten oder Zeugnisse;

- *verschiedenartige Gespräche, etwa Exploration oder Erhebung einer Anamnese;*
- *unterschiedliche Arbeitsproben, ja sogar, im Rahmen von Verhaltensbeobachtungen, unspezifische Verfahren.*

9.2 Analyse und Vergleich von Testdaten: Profilanalyse und Profilvergleich

Kapitel 9.2 bespricht zwei praktische Fragen, die Analyse und den Vergleich von Daten, die mit Leistungstests erhoben wurden. Besprochen werden

- die Profilanalyse (9.2.1),
 - und der Profilvergleich (9.2.2),
- Das Teilkapitel schließt mit einem Resümee (9.2.3).

Beide Aufgaben verschränken sich, sind aber je um einen eigenen Schwerpunkt zentriert.

9.2.1 Profilanalyse

Die Profilanalyse klärt das Anliegen, wie ein Untersucher Testergebnisse, die er erhoben hat, sinnvoll aufschlüsseln und welche Aussagen er aus den Testscores ableiten kann. Sie läßt sich in fünf Abschnitte gliedern:

1. Ermittlung der Rohpunkte,
2. Umwandlung der Rohpunkte in Standardwerte,
3. Klassifikation der Testwerte nach Normgruppen-Mittelwert,
4. Klassifikation der Testwerte nach individuellem Mittelwert,
5. Zusammenfassung der Analyse-Ergebnisse in einem Untersuchungsbericht.

Die fünf Schritte veranschaulichen wir an Daten, die mit dem „Leistungsprüfsystem“ von Horn gewonnen wurden (LPS, 1983). Dabei verwenden wir die Begriffe ‚Rohpunkte‘ und ‚Standardwerte‘; erklärt werden die Konzepte in dem Kapitel über ‚Normierung‘ (S. 111): *Rohpunkte* repräsentieren die Summe der Punkte, die je Richtiglösung gegeben werden. Transformiert man die Rohpunkte so, daß verschiedene Tests oder Subtests gleiche Mittelwerte und gleiche Standardabweichungen erhalten, so erstellt man *Standardwerte* - sie erleichtern den Vergleich von Testwerten.

1. Schritt: Erhebung der Rohpunkte

Herr Kleiber, 18 Jahre alt, habe das „Leistungs-Prüf-System (LPS)“ von Horn bearbeitet. Der Test ist nach dem „Gruppenfaktorenmodell“ von Thurstone konzipiert (S. 268); er besteht aus vierzehn Untertests, die sechs kognitive Teilfunktionen, sogenannte „Primärfaktoren“, erfassen.

Für eine richtige Lösung erhält der Proband einen Rohpunkt. Kasten 9-4 stellt die Subtests vor und gibt die Rohpunkte an, die Herr Kleiber erreicht hat.

Kasten 9-4:
Leistungs-Prüf-System (LPS): Rohpunkte in Untertests und im Gesamttest

Untertest-Nummer	Untertest-Name/Erfaßtes Merkmal	Rohpunkte
(1 + 2)	Allgemeinbildung/Wortverständnis	11
(3 + 4)	Denkfähigkeit	56
(5 + 6)	Wortflüssigkeit	43
(7 bis 10)	Räumliche Vorstellung	132
(11 + 12)	Gestaltbindung/Ratefähigkeit	33
(13 + 14)	Wahrnehmungstempo	52
GL	Gesamtintelligenz	327

2. Schritt: Umwandlung der Rohpunkte in Standardwerte

Die Rohpunkte werden in Standardwerte umgewandelt - nach der Normtabelle, welche die Handanweisung für die 18jährigen anführt (Horn, 1983, 22). Kasten 9-5 bringt in der dritten Spalte erneut die Rohpunkte, in der vierten die Standardwerte: in diesem Falle **Centilwerte** (C-Werte: der Mittelwert liegt bei 5, die Standardabweichung beträgt 2).

Für alle C-Werte sei auch der *Vertrauensbereich* (VB) mitbestimmt, der angibt, in welchem Intervall mit einer gewählten Restwahrscheinlichkeit der „wahre Wert“ liegt. Berechnung und Probleme bespricht Kapitel 4 (Reliabilität, S. 89).

Kasten 9-5:
Leistungs-Prüf-System (LPS) von Horn: Standardwerte (Centile)

C-Werte : Centile (sie reichen von -1 bis + 11)				
VB : Vertrauensbereich der C-Werte (ab-, aufgerundet; $p \leq 5\%$)				
GL : Gesamtleistung				
<i>Siehe auch den laufenden Text!</i>				
Untertest-Nummer	Untertest-Name/Erfaßtes Merkmal	Rohpunkte	C-Werte	VB
1 + 2	Allgemeinbildung/Wortverständnis	11	2	0.2-2.8
3 + 4	Denkfähigkeit	56	7	5.8-8.2
5 + 6	Wortflüssigkeit	43	5	4.4-5.6
7 bis 10	Räumliche Vorstellung	132	9	8.6-9.4
11 + 12	Gestaltbindung/Ratefähigkeit	33	3	2.2-3.8
13 + 14	Wahrnehmungstempo	52	8	7.6-8.4
GL	Gesamtintelligenz	327	6	5.6-6.4

Die Standardwerte erlauben verschiedene Klassifizierungen, nur zwei seien eingeführt:

- Klassifikation der Testwerte nach Normgruppen-Mittelwert: *Schritt 3,*
- Klassifikation der Testwerte nach individuellem Mittelwert: *Schritt 4.*

3. Schritt: Klassifikation der Testwerte nach Normgruppen-Mittelwert

Die Standardwerte (Centilwerte) von Herrn Kleiber in Kasten 9-4 seien verglichen mit den Standardwerten der 18jährigen, die zur Eichstichprobe gehören (LPS-Manual, 1983, 22). Bezugspunkt des Vergleiches ist der Mittelwert, der bei $C = 5$ liegt. Wir beschränken die Klassifikation auf drei Bereiche; dabei stehe M für Mittelwert, SD für Standardabweichung:

- Als *durchschnittlich* gelte eine Leistung in dem Bereich $M \pm 1 SD$, hier also in dem Intervall $C = 3$ bis $C = 7$.
- Als *überdurchschnittlich* gilt eine Leistung, die $C = 7$ übersteigt.
- Als *unterdurchschnittlich* gilt eine Leistung, die unter $C = 3$ liegt.

HINWEIS: Wenn Werte an der Grenze zweier Klassen liegen (also bei $C = 3$ und $C = 7$), sei die Lage eigens erwähnt. Warum? Punktuell genommen, gehören sowohl $C = 3$ als auch $C = 7$ zum Durchschnittsintervall. Beachtet man jedoch den Vertrauensbereich, dann kann der „wahre Wert“ für $C = 3$ im durchschnittlichen, aber auch im unterdurchschnittlichen Bereich, für $C = 7$ im durchschnittlichen, aber auch im überdurchschnittlichen Bereich liegen.

Nun die Klassifikation:

Im Vergleich zur Altersgruppe der 18jährigen lassen sich die Leistungen von Herrn Kleiber einstufen wie folgt (in Klammern die zugehörigen Testbelege):

Als *durchschnittlich* erwiesen sich

- die Gesamtintelligenz (GL),
- die Denkfähigkeit (3 + 4)
(Grenzbereich durchschnittlich/überdurchschnittlich),
- die Wortflüssigkeit (5 + 6),
- die Gestaltbindung (11 + 12)
(Grenzbereich durchschnittlich/unterdurchschnittlich).

Überdurchschnittlich waren

- das räumliche Vorstellen (7 bis 10),
- das Wahrnehmungstempo (13 + 14).

Unterdurchschnittlich war

- das Wortverständnis (1 + 2).

Was leisten die Vergleiche mit der Altersgruppe? Sie sagen etwas darüber aus, wie Herrn Kleibers Intelligenz - gemessen mit dem LPS - ausgeprägt ist bezüglich seiner Altersgenossen. Es handelt sich um eine stichprobenab-

1 Zur Klassifikation: Es gibt Autoren, welche die Einteilungsklassen anders abgrenzen als in unserem Beispiel. - Siehe Kap. 4 (Normierung, S. 118)!

hängige Feststellung (ähnlich wie wenn man sagt: Der 6jährige Günter mit seiner Größe von 106 cm gehört zu dem ‚unteren‘ Drittel seiner Altersgruppe).

4. Schritt: Klassifikation der Testwerte nach individuellem Mittelwert

Das Testprofil erlaubt zwei weitere Vergleiche:

1. einen Vergleich der Untertests mit dem individuellen Mittelwert,
2. einen Vergleich der Untertests miteinander.

Zu 1: Den individuellen Mittelwert soll hier der LPS-Gesamtwert (GL) repräsentieren; die Frage lautet:

- Steigt der Score eines Untertests erheblich über den Gesamtwert (GL)? Wenn ja, liegt ein *Leistungshoch* vor.
- Sinkt der Score eines Untertests erheblich unter den Gesamtwert (GL)? Wenn ja, liegt ein *Leistungstief* vor.

Zu 2: Bei einem Vergleich der Untertests miteinander lautet die Frage: Unterscheiden sich die Standardwerte **eines** Untertests erheblich von denen eines **anderen** Untertests?

In beiden Fällen ist zu fragen: Macht der Vergleich Sinn im Kontext einer konkreten Fragestellung?

Wir beschränken uns auf den Vergleich der Untertests mit dem individuellen Mittelwert, also mit dem LPS-Gesamtwert (GL). Wir fragen demnach nur nach den individuellen Leistungsschwerpunkten (Hochs) und den individuellen Leistungstiefpunkten (Tiefs).

Problem: Der Vergleich beruht auf der Berechnung der Kritischen Differenzen, schließt demnach auch ihre Problematik ein, die darin besteht, daß eine Stichprobenstatistik (hier die Reliabilität) übertragen wird auf einen Einzelfall (Kap. 4, S. 93).

Die **Kritische Differenz** (KD) berechnet sich - angewandt auf unser Beispiel - nach der Formel:

$$KD_{UT-GL} = z_{\alpha} \cdot SX \cdot \sqrt{2 - (r_{GL} + r_{UT})}$$

Es bedeuten:

KD_{UT-GL} : Kritische Differenz zwischen einzelnen Untertests (UT) und Gesamtleistung (GL),

z_{α} : gewählte Irrtumswahrscheinlichkeit, hier $p = 5\%$, demnach: $z_{\alpha} = 1.96$,

SX : Standardabweichung der verwandten Werte,
hier bei C-Werten: $SX = 2$,

r_{GL} : Reliabilität der Gesamtleistung (GL),

r_{UT} : Reliabilität der einzelnen Untertests (UT).

Die Kritische Differenz (KD) wird verglichen mit der Empirischen Differenz (ED).

Die **Empirische Differenz** (ED) ergibt sich aus der Subtraktion des C-Wertes von GL und des C-Wertes der jeweiligen Untertests: ($C_{UT} - C_{GL}$).

Zu **entscheiden** ist dann wie folgt:

- Ist die Empirische Differenz (Absolutbetrag) größer als die Kritische Differenz, dann besteht ein signifikanter Unterschied zwischen GL und UT.
- Ist die Empirische Differenz (Absolutbetrag) kleiner als die Kritische Differenz (oder ist sie gleich Null), dann besteht kein signifikanter Unterschied.

Kasten 9-6 bringt ein Beispiel.

Kasten 9-6:

Vergleich von Kritischer und Empirischer Differenz zur Ermittlung von Leistungshoch oder Leistungstief im Leistungs-Prüf-System (LPS)

Zwischen der ‚Gesamtleistung‘ (GL: C = 6) und dem Untertest (1 + 2: Wortverständnis: C = 2) werden Empirische und Kritische Differenz ermittelt.

Empirische Differenz (ED): **ED = 2 C - 6 C = -4 C = 1 4 |**

Kritische Differenz (KD): **KD_{GL - (1+2)} = 0.87 C**

Daten zur Berechnung der Kritischen Differenz:

$r_{it, GL} = 0.99$ $r_{it, (1+2)} = 0.96$; $p = 5 \% \rightarrow z = 1.96$

$$KD_{GL - (1+2)} = 1.96 \cdot 2 \cdot \sqrt{2 - (0.99 + 0.96)} = 0.87$$

Entscheidung: Da ED (mit C = 1 4 |) größer ist als KD (mit C = 0.87), weicht der C-Wert von Untertest (1+2) signifikant vom dem C-Wert der Gesamtleistung (GL) ab; er liegt erheblich unterhalb des individuellen Mittelwertes. - In „Wortverständnis“ (Untertest 1+2) zeigt sich demnach ein *individuelles Leistungstief* an.

Kasten 9-7:

Kritische Differenzen zwischen dem C-Wert der Gesamtleistung (GL) und den C-Werten der einzelnen Untertests (UT) im Leistungs-Prüf-System (LPS) bei dem Probanden Herrn Kleiber

UT : Untertest

ED : Empirische Differenz ($C_{UT} - C_{GL}$)

KD : Kritische Differenz

Hoch : Leistungshoch

Tief : Leistungstief

NS : Nicht signifikant (Weder Hoch noch Tief)

UT	r_{it}	C-Wert	ED	KD	Ergebnisse		
					Hoch	Tief	NS
(1 + 2)	0.96	2	- 4	0.87		X	
(3 + 4)	0.90	7	+1	1.30			X
(5 + 6)	0.98	1	- 5	0.67		X	
(7 bis 10)	0.99	9	+3	0.55	X		
(11 + 12)	0.96	3	- 3	0.87		X	
(13 + 14)	0.99	8	+2	0.55	X		
GL	0.99	6					

Kasten 9-7 referiert die Ergebnisse eines Vergleiches aller Untertests mit der Gesamtleistung (GL), dem Repräsentanten des individuellen Mittelwertes.

Was Kasten 9-7 in Zahlen darstellt, sei nun in Worten formuliert (Testbelege in Klammern):

In dem Testprofil fanden sich

- *Leistungsschwerpunkte/Leistungshochs*
 - im räumlichem Vorstellen (7 bis IO),
 - im Wahrnehmungstempo (13 + 14).
- *Leistungstiefpunkte/Leistungstiefs*
 - im Wortverständnis (1 + 2),
 - in Wortflüssigkeit (5 + 6),
 - in Gestaltbindung (11 + 12).

Wo führt die Ermittlung ‚individueller Hochs und Tiefs‘ über die Vergleiche mit der Altersgruppe hinaus? Die Aussagen führen insofern weiter, als sie den Probanden in seiner individuellen kognitiven Struktur charakterisieren. Der Untersucher kann, in unserem Beispiel, *Herrn Kleiber in seinen kognitiven Stärken und Schwächen kennzeichnen* - unabhängig von einem Vergleich mit den Gleichaltrigen.

5. Schritt: Zusammenfassung der Analyse-Ergebnisse in einem Untersuchungsbericht

Adressat einer Profilanalyse ist in der Regel ein Klient, der sich an den Untersucher gewandt hat. Dieser Adressat dürfte dankbar sein, wenn ihm die Ergebnisse übersichtlich dargeboten werden. Eine solche Zusammenfassung erhält hier den Titel ‚Untersuchungsbericht‘.

Mit dem ‚Untersuchungsbericht‘ wählen wir eine bestimmte Art der Zusammenfassung. Sie kann knapp oder ausführlich gehalten sein, sie kann mündlich mitgeteilt oder schriftlich gefaßt werden. - Im Rahmen eines Gutachtens hat der ‚Untersuchungsbericht‘ eine genau umschriebene Funktion (Kap. 21, S. 447). Hier seien die Analyse-Ergebnisse ohne Kommentar zu einem solchen Bericht zusammengestellt.

Untersuchungsbericht

Testbeschreibung: Das Leistungs-Prüf-System von Horn (LPS) soll wichtige Grundfähigkeiten der Intelligenz erfassen. Repräsentiert werden diese Dimensionen in unterschiedlichen Aufgabengruppen, die zu Untertesteinheiten zusammengefaßt werden. Normen liegen für verschiedene Altersgruppen vor.

Verhaltensbeobachtung: Herr Kleiber arbeitete bereitwillig mit . . . usw.

Ergebnisbericht: Herr Kleiber erzielte in den einzelnen Untertests (UT) folgende *Standardwerte*, mitgeteilt in Centilen (C: Mittelwert = 5, Standardabweichung = 2). Die Werte beziehen sich auf die Gruppe der 18jährigen. Für die C-Werte sei der *Vertrauensbereich* (VB) mitbestimmt.

Die *Merkmalsausprägung* wird nur angegeben in den drei Klassen: durchschnittlich (d), unterdurchschnittlich (u-d) und überdurchschnittlich (ü-d). Wenn jedoch Werte an der Grenze zweier Klassen liegen (bei C = 3 und C = 7), sei die Lage miterwähnt*.

UT	Erfaßtes Merkmal	C	VB	Ausprägung
1 + 2	Allgemeinbildung/Wortverständnis	2	0.2-2.8	u - d
3 + 4	Denkfähigkeit	7	5.8-8.2	d (Grenze ü-d)
5 + 6	Wortflüssigkeit	5	4.4-5.6	d
7 bis 10	Räumliche Vorstellung	9	8.6-9.4	ü - d
11 + 12	Gestaltbindung/Ratefähigkeit	3	2.2-3.8	d (Grenze u-d)
13 + 14	Wahrnehmungstempo	8	7.6-8.4	ü - d
GL	Gesamtintelligenz	6	5.6-6.4	d

Faßt man die Leistungen über alle Untertests zu einer sogenannten *Gesamtleistung* (GL) zusammen, so ergibt sich ein C-Wert von 6, das entspricht einem Prozentrang von 69. (Der Prozentrang von 69 besagt, daß rund dreißig Prozent der Altersgruppe höhere Leistungen erbringen.) - Geht man bei dieser *Gesamtleistung* von einer Irrtumswahrscheinlichkeit von fünf Prozent aus, so liegt der wahre Wert in dem Intervall von C = 5.6 bis C = 6.4.

Interpretation: Die Gesamtintelligenz ließ sich als durchschnittlich einstufen (im oberen Durchschnittsbereich).

Im Ergleichen zur Altersgruppe gilt:

- **Überdurchschnittlich** ausgeprägt waren
 - das räumliche Vorstellen und
 - das Wahrnehmungstempo.
- Als **durchschnittlich** erwiesen sich
 - die Denkfähigkeit (Grenze durchschnittlich/überdurchschnittlich)
 - die Wortflüssigkeit und
 - die Gestaltbindung (Grenze durchschnittlich/unterdurchschnittlich).
- **Unterdurchschnittlich** war die Leistung
 - im Wortverständnis.

Im Vergleich zur eigenen mittleren Leistung fanden sich

- **Leistungshochs**

2 Die Werte, die hier an der Grenze zweier Klassen liegen (C = 3 und C = 7), gehören zum Durchschnittsintervall. Beachtet man jedoch den Vertrauensbereich, dann kann der „wahre Wert“ für C = 3 im durchschnittlichen, aber auch im unterdurchschnittlichen Bereich, für C = 7 im durchschnittlichen, aber auch im überdurchschnittlichen Bereich liegen.

- im räumlichen Vorstellen und
- im Wahrnehmungstempo.
- **Leistungstiefs** dagegen ergaben sich
 - im Wortverständnis,
 - in Wortflüssigkeit,
 - in Gestaltbindung.

9.2.2 Profilvergleich

Der Profilvergleich hat das Ziel, ein gegebenes Leistungsprofil (von Individuum oder Gruppe) mit einem anderen Profil zu vergleichen. Beispielsweise läßt sich das Profil eines Berufsanwärters vergleichen mit dem Profil erfolgreicher Berufsvertreter.

Der Berufseignungstest von Schmale und Schmittke (BET 1966) bietet diese Möglichkeit an; der Tabellenband legt Berufsprofile vor, mit denen das Profil eines Probanden verglichen werden kann (1966, S.. 36). Die Frage lautet: Ist das Profil des Berufsanwärters dem Profil der Berufsvertreter so ähnlich, daß ihm geraten werden kann, den Beruf zu wählen?

Der Profilvergleich ist auf vielfältige Weise möglich. Es seien fünf Wege benannt:

1. Korrelation,
2. Distanzmaß nach Osgood und Suci,
3. Kombination von Korrelation und Distanzmaß nach Cattell,
4. Distanzmaß nach Kristof,
5. Ähnlichkeitsmaß nach Huber, H.P.

Die fünf Möglichkeiten seien mit denselben Profilen durchgespielt, damit die Unterschiede deutlich hervortreten. Wir verwenden -wie im vorigen Abschnitt bei der Profilanalyse - Daten zum ‚Leistungsprüfsystem‘ von Horn (LPS: 1983). Wir geben zwei Profile vor:

- ein Gruppenprofil und
- ein individuelles Probandenprofil.

Das **Gruppenprofil** dient als Kriterium, mit dem das **Probandenprofil** verglichen wird. Kasten 9-8 führt beide Profile an.

Kasten 9-8:

Leistungsprüfsystem (LPS): Ein Gruppen- und ein Probandenprofil: Vergleichsdaten

Centil-Werte in einem LPS-Profil/UT: Untertest

U T .	1 + 2	3 + 4	5 + 6	7 bis 10	11 + 12	13 + 14	GL
Gruppenprofil	2	5	8	4	6	9	5.7
Probandenprofil	4	6	9	5	7	8	6.5

(1) Korrelation

Beantwortet werden soll die Frage: Wie ähnlich ist das Profil des Probanden dem Gruppenprofil?

Als Koeffizienten wählen wir das nonparametrische ‚rho‘ (ρ) - dessen Berechnung auf Rangreihen beruht. Kasten 9-9 führt die Rangreihen der beiden Profile an.

Kasten 9-9:

Gruppenprofil und Probandenprofil: Rangreihen der beiden Profile aus Kasten 9-8

UT:	Untertest						
Gruppenprofil:	Rangreihe 1						
Probandenprofil:	Rangreihe 2						
„d“:	Differenz zwischen Rangreihe 1 und 2						
<i>Siehe die Veranschaulichung unterhalb des Kastens!</i>							
	U T ▶ 1 + 2	3+4	5+6	7 bis 10	11 + 12	13 + 14	GL
Gruppenprofil	1	3	6	2	5	7	4
Probandenprofil	1	3	1	2	5	6	4
„d“	0	0	-1	0	0	1	0

Der Koeffizient „ ρ “ berechnet sich nach der Formel:

$$\rho = 1 - \frac{6 \sum d^2}{N(N^2 - 1)}$$

Es bedeuten:

N : Zahl der Meßwerte, hier der 7 Untertests,

d : Differenz zwischen den Rangreihen des Gruppen- und des Probandenprofils.

Veranschaulichung - Ermittlung der Differenzen „d“:

- Im **Gruppenprofil** liegt der niedrigste Meßwert bei C = 2, der höchste bei C = 9. Da sieben Meßwerte vorliegen, sind sieben Rangplätze zu vergeben. Der niedrigste Wert erhält den Rangplatz 1, der höchste den Rangplatz 7.
- Im **Probandenprofil** liegt der niedrigste Meßwert bei C = 4, der höchste bei C = 9. Der C-Wert 4 erhält den Rangplatz 1, der C-Wert 9 den Rangplatz 7.
- Die beiden Rangreihen werden in Kasten 9-9 untereinander geschrieben, dann die Differenz „d“ zwischen den Rangplätzen gebildet.

Der Vergleich von Gruppen- und Probandenprofil ergibt:

$$\rho_{\text{Gruppe, Proband}} = 1 - \frac{6 \cdot 2}{7 \cdot (7^2 - 1)} = 0.96$$

Nach diesem Koeffizienten ist das Profil des Probanden dem Gruppenprofil sehr ähnlich.

Einwand: Gegen die Prozedur, wie eben beschrieben, spricht die Tatsache, daß ein Korrelationskoeffizient nur die Gleichheit der Variation zweier Meßwertreihen „abbildet“; unbeachtet bleibt, wie weit die beiden Meßwertreihen voneinander entfernt verlaufen. Verliefe etwa die Meßwertreihe des Probanden je Untertest um zwei C-Werte weiter entfernt von dem Gruppenprofil: Der Koeffizient bliebe bei $\rho = 0.96$.

Darum eignen sich Korrelationskoeffizienten nur in begrenztem Maße zum Profilvergleich.

(2) Distanzmaß von Osgood und Suci

Was der Korrelationskoeffizient nicht beachtet, nimmt der Vorschlag von Osgood und Suci eigens auf (1952): In den Profilvergleich gehen die Distanzen ein, und zwar nach der Formel:

$$D = \sqrt{\sum d^2}$$

Es bedeuten:

D : Distanzmaß nach Osgood und Suci,

d : Differenz zwischen den entsprechenden Test-Scores von Gruppenprofil (Gr) und Probandenprofil (Pb), hier zwischen den Centilwerten.

Beispiel: $d_l = Gr_{Subtest (1+2)} - Pb_{Subtest (1+2)} = 2 C - 4 C = - 2 C$.

Kasten 9-10 veranschaulicht die Prozedur an den Werten von Kasten 9-8.

Kasten 9-10:
Gruppenprofil und Probandenprofil:
Differenzen zwischen den Centilwerten der beiden Profile aus Kasten 9-8

UT : Untertest							
„d“ : Differenz zwischen C-Werten des Gruppen- und des Probandenprofils							
UT ►	1 + 2	3 + 4	5 + 6	7 bis 10	11 + 12	13 + 14	GL
Gruppenprofil	2	5	8	4	6	9	5.7
Probandenprofil	4	6	9	5	7	8	6.5
„d“	-2	-1	-1	-1	-1	1	-0.8

Das Osgood-Suci-Distanzmaß beträgt:

$$D = \sqrt{9.64} = 3.10$$

Einwund: Das Distanzmaß von Osgood und Suci bleibt vergleichsweise unanschaulich; es gibt nur ein Mehr oder Weniger an. - Läßt sich die Ähnlichkeit von Profilen nicht anschaulicher darstellen? Einem solchen Anliegen dient das Maß, das Cattell vorgeschlagen hat.

(3) Ähnlichkeitsindex von Cattell

Cattell wandelt das Distanzmaß von Osgood und Suci in einen Koeffizienten um, der sich analog einem Korrelationskoeffizienten interpretieren läßt.

Blicken wir kurz zurück: Vorschlag 1 (Korrelationskoeffizient) ist anschaulich, beachtet aber die Distanz zweier Meßwertreihen nicht. Vorschlag 2 (Distanzmaß) beachtet die Distanz zweier Meßwertreihen, bleibt aber unanschaulich. Vorschlag 3 (Ähnlichkeitsindex) kombiniert beide Anliegen: Distanzmaß und Analogie zu einem Korrelationskoeffizienten.

Der Vorschlag von Cattell lautet (1949, 1969):

$$r_{\text{profil}} = \frac{2 \cdot X^2_{0.50(FG)} \cdot s^2 - D^2}{2 \cdot X^2_{0.50(FG)} \cdot s^2 - D^2}$$

Es bedeuten:

- r_{profil} : Index für die Ähnlichkeit zweier Profile (interpretierbar wie ein Korrelationskoeffizient),
- $\chi^2_{0.50(FG)}$: Chi-Quadrat-Wert für $p = 50$ bei den entsprechenden Freiheitsgraden,
- FG : Freiheitsgrade, hier Zahl der verglichenen Tests, also $FG = 7$,
- s^2 : Standardabweichung der Profilwerte, hier bei C-Werten: $s^2 = 2^2 = 4$;
- D^2 : Distanzmaß nach Osgood-Suci (wie unter Nummer 2 vorgestellt).

Die Quadratsummen der Differenzen (D^2) folgen einer Chi-Quadrat-Verteilung.

Für den Vergleich ergibt sich:

$$r_{\text{profil}} = \frac{2 \cdot 6.35 \cdot 2^2 - 9.64}{2 \cdot 6.35 \cdot 2^2 + 9.64} = \frac{41.16}{60.44} = 0.68$$

Einwand: Der Ähnlichkeitsindex ist ‚anschaulich‘ wie ein üblicher Korrelationskoeffizient, erlaubt jedoch keine signifikanz-statistische Entscheidung über die Ähnlichkeit der verglichenen Profile.

Dies leistet die nächste Prozedur.

(4) Distanzmaß nach Kristof

Kristof bezieht in den Profilvergleich die Reliabilität der Tests mit ein (1957). Damit stellt sich erneut die Frage, wie weit Gruppenstatistiken (hier die Reliabilität) übertragen werden dürfen auf Einzelfälle oder kleinere Gruppen (Kap. 5, Reliabilität, S. 93).

Kristof spricht mehrere Arten von Vergleichen durch. Hier sei nur der Fall angeführt, bei dem ein *Individualprofil* mit einem Gruppenprofil verglichen wird; die Formel lautet:

$$X^2 = \frac{N}{(N+1) \cdot s^2} \Sigma \frac{D^2}{1 - r_{tt}}$$

Es bedeuten:

N : Umfang der Stichprobe, an der das Profil gewonnen wurde;
hier sei N = 115;

D² : Quadratsummen der Distanzmaße nach Osgood-Suci
(wie unter Nummer 2 vorgestellt).

Kasten 9-11 veranschaulicht den Vergleich zwischen Gruppen- und Probandenprofil.

Kasten 9-11:

Vergleich eines Individualprofils mit einem Gruppenprofil nach Kristof (1957)

UT: Untertest Gruppen- und Probandenprofil geben C-Werte vor.							
<i>Siehe den laufenden Text!</i>							
UT ►	1 + 2	3 + 4	5+6	7 bis 10	11 + 12	13 + 14	GL
Gruppenprofil	2	5	8	4	6	9	5.7
Probandenprofil	4	6	9	5	7	8	6.5
D ²	4	1	1	1	1	1	0.64
r _{tt}	0.96	0.90	0.98	0.99	0.96	0.99	0.99
1 - r _{tt}	0.04	0.10	0.02	0.01	0.04	0.01	0.01
D ² /(1-r _{tt})	100	10	50	100	25	100	64

Für den Vergleich ergibt sich:

$$\frac{N}{(N+1) s^2} = \frac{115}{(115+1) 2^2} = 0.2478$$

$$\Sigma \frac{D^2}{1 - r_{tt}} = (100 + 10 + 50 + 100 + 25 + 100 + 64) = 449$$

Einsetzen in die Formel von Kristof ergibt den *empirischen* X²-Wert:

$$X^2 = 0.2478 \cdot 449 = 111.28$$

Der *kritische* Wert für X² bei 7 Freiheitsgraden (hier: 7 Untertests) und einer Restwahrscheinlichkeit von p 15 % beträgt X² = 14.067.

Entscheidung: Der *empirische* X²-Wert ist erheblich größer als der *kritische* X²-Wert. Demnach muß (mit p ≤ 5 %) die Hypothese der Profilähnlichkeit zwischen Gruppen- und Probandenprofil verworfen werden.

Wenn die Reliabilität in den Profilvergleich eingeht (unter den gleichen Vorbehalten wie bei Berechnung Kritischer Differenzen), dann ergibt sich in unserem Beispiel eine Unähnlichkeit der verglichenen Profile.

(5) Ähnlichkeitsmaß nach Huber

Huber, H.P. verbindet die Konzepte von Cattell und Kristof: Wie Cattell entwickelt er einen Index analog einem Korrelationskoeffizienten; wie Kristof nimmt er die Reliabilität in die Berechnung mit auf.

Es sei nur die generelle Formel angeführt (1973 b, 41):

$$r_p = \frac{X^2_{0.50(FG)} \cdot X^2}{X^2_{0.50(FG)} + X^2}$$

Es bedeuten:

- r_p : Profilähnlichkeitskoeffizient
 $X^2_{0.50(FG)}$: Chi-Quadrat-Wert für $p = 50$
 bei den entsprechenden Freiheitsgraden (FG),
 FG : Freiheitsgrade, hier Zahl der verglichenen Tests, also $FG = 7$,
 X^2 : X^2 -Wert, wie ihn der Vergleich bei Kristof ergeben hat.

Einsetzen der Werte erbringt das Ergebnis:

$$r_p = \frac{6.35 - 111.28}{6.35 + 111.28} = -0.89$$

Kasten 9-12:

Ausführliche Berechnung des Ähnlichkeitsmaßes nach Huber, H.P. (1973 b, 43)

$$r_{global} = \frac{(1 + N) \cdot s^2 \cdot X^2_{0.50(FG)} - N \cdot \sum \frac{D^2}{1 - r_{it}}}{(1 + N) \cdot s^2 \cdot X^2_{0.50(FG)} + N \cdot \sum \frac{D^2}{1 - r_{it}}}$$

Es bedeuten:

- r_{global} : Globaler Profilähnlichkeitskoeffizient,
 N : Umfang der Stichprobe, an der das Profil gewonnen wurde, hier sei $N = 115$,
 s^2 : Standardabweichung der Profilwerte, hier bei C-Werten: $s^2 = 2^2$,
 $X^2_{0.50(FG)}$: Chi-Quadrat-Wert für $p = 50\%$ bei den entsprechenden Freiheitsgraden (FG),
 hier: $FG = 7 \Rightarrow X^2 = 6.35$,
 $\sum \frac{D^2}{1 - r_{it}}$: Summe der letzten Zeile des Kastens 9-11, der Betrag lautet: 449.
Alle Werte sind bekannt. Einsetzen ergibt:

$$r_{global} = \frac{(1 + 115) \cdot 2^2 \cdot 6.35 - 115 \cdot 449}{(1 + 115) \cdot 2^2 \cdot 6.35 + 115 \cdot 449} = -0.89$$

Die ausführliche Berechnung in diesem Kasten 9-12 erbringt dasselbe Ergebnis wie die Berechnung nach der generellen Formel zu Beginn des Abschnitts 5.

Entscheidung: Nach dem Ähnlichkeitsmaß stehen Gruppen- und Probanden-Profil in einem engen, aber gegenläufigen Zusammenhang.

HINWEIS: Für eine Berechnung unmittelbar aus den Daten gibt Huber, H. P. eine eigene Formel an (1973 b, 43). Kasten 9-12 führt die Formel an.

9.2.3 Resümee zu Kapitel 9.2

Das Teilkapitel 9.2 beschreibt zwei praktische Aufgaben, die sich stellen können, wenn ein Untersucher Testwerte erhoben hat.

Kapitel 9.2.1 schildert unterschiedliche Schritte einer Analyse dieser Testwerte. Sie betreffen die Ermittlung der Rohpunkte, ihre Umwandlung in Standardwerte, unterschiedliche Vergleiche und ihre Zusammenfassung in einem Untersuchungsbericht.

Kapitel 9.2.2 schildert unterschiedliche Möglichkeiten, Testwerte eines Individuums (oder einer Gruppe) zu vergleichen mit den Testwerten eines anderen Individuums oder einer Gruppe.

9.3 Kriterien und Beispiel einer Testbewertung

Das Teilkapitel 9.3

- referiert allgemeine Kriterien einer Testbewertung (9.3.1),
 - und gibt ein Einzelbeispiel für die Bewertung (9.3.2).
- Es folgt ein Resümee (9.3.3).

9.3.1 Kriterien einer Testbewertung

Ob ein Test zu verwenden ist, sollte nicht seine psychometrische Qualität allein bestimmen. In diese Entscheidung sollte vor allem auch die Einschätzung des Beitrags eingehen, den er zur Beantwortung der konkreten Fragestellung verspricht (Michel & Conrad, 1982, 64-65; siehe Kap. 17, Nutzenschätzung s. 373).

Diese Beurteilung erfordert, daß der Diagnostiker in der Lage ist, Tests zu bewerten. Um Bewertungen zu erleichtern, hat das Testkuratorium **der Förderung Deutscher Psychologenvereinigungen** Kriterien veröffentlicht (1986). Die Kriterien seien vollständig zitiert, sie gelten nicht nur für Leistungs-, sondern auch für Persönlichkeitstests.

Beschreibung der einzelnen Kriterien für die Testbeurteilung

Präambel

Im folgenden werden einige Gesichtspunkte angegeben, die bei der Beurteilung eines Testverfahrens von Bedeutung sind. Teilweise handelt es sich bei den genannten Aspekten um widersprüchliche Forderungen, denen kaum jemals von einem Testverfahren gleichermaßen Rechnung getragen werden kann. Bei der Beurteilung einer diagnostischen Methode kommt es auf die jeweils besonderen Umstände, Bedingungen und Zielsetzungen an, die aber deutlich zu erkennen und nachvollziehbar sein müssen; vor diesem Hintergrund werden die genannten Kriterien im Einzelfall zu gewichten sein.

Grundlage für die Testbewertung ist prinzipiell das Testmanual; dieses muß so beschaffen sein, daß die wichtigsten Aussagen zu den für die Beurteilung relevanten Punkten daraus erarbeitet werden können.

1. Testgrundlage

1.1. Diagnostische Zielsetzung. Die Angaben zu diesem Punkt sollen es dem Benutzer ermöglichen, den Beitrag des Verfahrens zu einer diagnostischen Entscheidungsfindung zu erkennen. Dies betrifft sowohl den prinzipiellen diagnostischen Ansatz (etwa Zustandsdiagnostik, Veränderungsmessungen) als auch den vom Testautor intendierten Beitrag im Rahmen einer umfassenderen diagnostischen Informationssammlung. Wenn das Verfahren von seiner Zielsetzung oder vom aktuellen Entwicklungsstand her nicht für eine Einzelfalldiagnostik geeignet, sondern nur für Forschungszwecke vorgesehen ist, sollte dies explizit angegeben werden.

1.2. Theoretische Grundlagen. Hier ist deutlich zu machen, in welcher Weise das Verfahren auf den Ergebnissen der wissenschaftlichen Psychologie aufbaut. Die relevante Grundkonzeption muß ohne zusätzliche Sekundärliteratur erkennbar sein. Modifikationen etablierter theoretischer Vorstellungen durch den Testautor sind besonders deutlich zu machen.

1.3. Nachvollziehbarkeit der Testkonstruktion. Der Benutzer muß durch die Angaben zu diesem Punkt in angemessener ausführlicher und verständlicher Weise in die Lage versetzt werden, die einzelnen Arbeitsschritte bei der Erstellung der Testmaterialien kritisch zu bewerten. Hierzu gehören insbesondere Angaben über die Veränderung bzw Selektion von ursprünglich aus theoretischen Überlegungen heraus zusammengestellten Indikatorenmengen.

2. Testdurchführung

2.1. Durchführungsobjektivität. Das Ausmaß, in dem die Unabhängigkeit des Tests von der Person des Untersuchungsleiters durch die Vorschriften der Testinstruktion und aller übrigen Durchführungsbedingungen gesichert ist.

2.2. Transparenz Das Ausmaß, in dem aus der Beschaffenheit eines Verfahrens die Spezifität und dessen Meßfunktion und Auswertung ersichtlich sind:

2.3. Zumutbarkeit Das Ausmaß, in dem ein Test (absolut und relativ zu dem aus der Anwendung des Verfahrens resultierenden Nutzen) die getestete Person in zeitlicher psychischer (insbesondere „energetisch“-motivational und emotional) sowie körperlicher Hinsicht beansprucht.

2.4. Verfälschbarkeit. Das Ausmaß, in dem ein Test die individuelle Kontrolle über Art und Inhalt der verlangten bzw. gelieferten Informationen ermöglicht.

2.5. Störanfälligkeit. Das Ausmaß, in dem ein Test zur Erfassung habituellem Merkmalsunterschiede unempfindlich gegenüber aktuellen Zuständen der Person und situativer Faktoren der Umgebung ist.

3. Testverwertung

3.1. Auswertungsobjektivität. Das Ausmaß, in dem die Auswertung des Tests unabhängig von personenbedingten oder apparativen Störquellen ist.

3.2. Zuverlässigkeit. Meßgenauigkeit oder Grad der Erklärbarkeit der beobachteten interindividuellen Unterschiede der Testergebnisse durch tatsächliche psychische Merkmalsunterschiede, untersucht etwa als Stabilität, Äquivalenz oder interne Konsistenz. Für die Bewertung ist die Angabe der verwendeten Berechnungsverfahren erforderlich.

3.3. Gültigkeit. Das Ausmaß der Treffsicherheit oder diagnostischen Valenz, mit dem der Test Rückschlüsse auf Verhalten außerhalb der Testsituation oder auf den Ausprägungsgrad des dem Testverhalten zugrundeliegenden Konstruktes ermöglicht. Bei der Testbeurteilung ist besonderes Schwergewicht auf die Ergebnisse zum Bereich der Kriteriumsvalidität zu legen.

3.4. Normierung. Ausmaß und Qualität der populationsspezifischen Bezugsgrößen zur Interpretation von Personenparametern, insbesondere zur Bestimmung der relativen Position einer Testperson in bezug auf (u.U. verschiedene) Populationsverteilungen von Testwerten.

3.5. Bandbreite. Ausmaß der Enge oder Vielfalt des Verfahrens gegenüber unterschiedlichen Fragestellungen, Gruppen- und Prognosezeiträumen.

3.6. Informationsausschöpfung. Menge und Qualität der Indikatoren, die bezogen auf verschiedene Ziele, Anlässe oder Probandengruppen begründet aus den Testantworten abgeleitet werden.

3.7. Änderungssensitivität. Möglichkeiten und Grade der Veränderungsmessung durch dieses Verfahren, insbesondere im Rahmen von Zeitreihenvergleichen.

4. Testevaluation

4.1. Ökonomie. Das Ausmaß, in dem ein Test bei der Durchführung, Auswertung und Anschaffung als kostengünstig zu bewerten ist.

4.2. Fairneß. Ausmaß einer eventuell bestehenden systematischen Diskriminierung bestimmter Testpersonen, z. B. aufgrund ihrer ethischen, soziokulturellen oder geschlechtsspezifischen Gruppenzugehörigkeit, bei der Abschätzung von Kriteriumswerten.

4.3. Akzeptanz. Ausmaß, in dem subjektive Meinungen, Bewertungen oder gesellschaftspolitische Überzeugungen gegen einen Test angeführt werden.

4.4. Vergleichbarkeit. Ausmaß der partiellen Übereinstimmung mit anderen Untersuchungsverfahren sowie die abweichenden Aspekte, Schwerpunkte oder Inhalte. Eine vermutete Sonderstellung bzw. Novität des Verfahrens ist besonders herauszustellen.

4.5. Bewährung. Systematische Aufarbeitung und Bewertung der mit dem Test gesammelten Erfahrungen, z.B. bezogen auf bestimmte Personengruppen oder diagnostische Ziele.

5. Äußere Testgestaltung

Die Verständlichkeit des Testmanuals, die probandenfreundliche Gestaltung der übrigen Testmaterialien sowie die Übereinstimmung von Titel und werblicher Darstellung mit dem tatsächlichen Testinhalt werden bei der Beurteilung herangezogen.

Stand: 09. Mai 1986“

9.3.2 Beispiel für eine Test-Bewertung (Fay, 1993)

Als Beispiel für eine Testbewertung, die sich an den Kriterien des Testkuratoriums orientiert, bringen wir die Besprechung, die Fay dem „HAWIE-R“ gewidmet hat (1993). Leider können wir nur Auszüge dieser auch sprachlich vorzüglichen Rezension wiedergeben; als Kernstück wählen wir den Abschnitt ‚Kritik‘ aus³.

Es geht um „Tewes, U. (Hrsg.). (1991). HAWIE-R: Hamburg- Wechsler Intelligenztest für Erwachsene, Revision 1991. Bern: Huber“.

Vorbemerkung: Ausführlich wird ein Text referiert, der auf viele Einzelheiten detailliert eingeht. Innerhalb dieses Buches ist diese differenzierte Art der Darstellung eine Ausnahme. Ich halte die Besprechung von Fay aber für das Paradigma einer fairen kritischen Stellungnahme zu einem oft verwandten Diagnosticum. Darum erscheint mir die Ausnahme berechtigt.

³ Herrn Dr. Ernst Fay danke ich für die Zustimmung, seinen Text in dieses Buch aufzunehmen

„1. **Testart:** Intelligenztest.

2. **Testmaterial:** Handbuch und Testanweisung (121 Seiten), Schachtel mit Testmaterial und achtseitigen Protokollbögen, zusätzlich: Stift und Stoppuhr.

3. **Testgliederung:** Der HAWIE-R besteht, wie sein Vorgänger, aus elf Skalen, die einem Verbalteil und einem Handlungsteil zugeordnet sind...

4. **Grundkonzept:** Dem Test liegt die Theorie der Intelligenz nach Wechsler zugrunde, die im Handbuch nicht dargestellt wird...

5. **Durchführung:** Die Anweisungen für die Durchführung des Tests sind in übersichtlicher Form dargeboten; der erheblich verbesserte Protokollbogen unterstützt den Testleiter...

6. **Auswertung:** In der Regel findet der Anwender ausreichend Hilfestellung in Form von Beispielen und übersichtlichen Darstellungen im Handbuch...

7. Gütekriterien

7.1 **Objektivität:** Von der Sicherstellung der Durchführungsobjektivität kann im allgemeinen ausgegangen werden...

7.2 **Reliabilität:** Von Ausnahmen abgesehen, können die im Handbuch mitgeteilten Reliabilitätskoeffizienten - es handelt sich ausschließlich um Maße der internen Konsistenz (Cronbachs Alpha), Untersuchungen zur zeitlichen Stabilität fehlen noch - befriedigen...

7.3 Validität:

Das Handbuch enthält . . . auf den knapp 17 Seiten zum Thema Validität lediglich

- 10 Seiten altersgruppenspezifische Interkorrelationsmatrizen,
- eine Seite Tabellen zu einer Faktorenanalyse (mit der klassischen 2-Faktoren-Lösung und 56 Prozent aufgeklärter Varianz),
- eine Seite mit einer tabellarischen Übersicht über Interkorrelationen der Testwerte von Familienangehörigen,
- eine Seite Daten zu Testergebnissen von Alkoholikern und einer Kontrollgruppe, und
- auf weiteren anderthalb Seiten sind die bereits in Tabellenform vorliegenden Daten zu altersspezifischen Leistungen in Form von Graphen noch einmal dargestellt.

Es bleiben 68 Zeilen Text zum Zentralkriterium eines jeden diagnostischen Verfahrens, zur Validität.

Dies ist kein Vorwurf an den Herausgeber, dies ist ein Aufruf an den Praktiker, den Test anzunehmen und sich selbst auf den mühsamen Weg der Validierung zu begeben.

7.4 **Normen:** Die Normierung erfolgte an einer hinsichtlich Schulbildung, Alter und Geschlecht für die Bundesrepublik des Jahres 1986 repräsentativen

Stichprobe (N = 2000)... Ganz offensichtlich konnte jedoch noch kein Einwohner der neuen Bundesländer Eingang in die Normierungsstichprobe finden, es ist schade, daß das im deutschsprachigen Bereich 'jüngstnormierte' Intelligenztestverfahren dort nur schwer zu interpretierende Ergebnisse zeitigen wird.

8. Kritik: Jeder Psychologe und jede Psychologin kennt ihn, und jede Psychologin und jeder Psychologe wußte es: Der Hamburg-Wechsler-Intelligenz-Test für Erwachsene, der HAWIE, war revisionsbedürftig... Es gab HAWIE-interne Probleme... Es gab HAWIE-externe-Probleme, denen (d.V.) der Autor mit der einleitenden salvatorischen Klausel begegnet: „Der amerikanische Herausgeber bestand auf eine möglichst enge Anlehnung der deutschen Version an das amerikanische Original“ (Handbuch, S. 1).

Dennoch stellt sich der Rezensent ebenso wie der Testbenutzer Fragen, die folgenden beispielsweise:

- „Kommt die Revision eines Testverfahrens ohne Nachdenken über eine Revision der Grundkonzeption aus, darf sie auskommen?“
- „Kann man, soll man, darf man anno 1991 eine 1939 entwickelte Intelligenzkonzeption einfach so übernehmen?“
- „Kann man so tun, als gäbe es die Kontroverse zwischen dem kognitionspsychologischen und dem psychometrischen Zugang zur Intelligenzerfassung gar nicht?“

Jetzt gibt es HAWIE-R Probleme

Problematisch erscheinen mir nach wie vor einige Testaufgaben - oder die Auswertungskriterien, je nach Sichtweise. Ich stelle diese Bedenken - zusammen mit einigen anderen - exemplarisch an einigen Items aus verschiedenen Untertests dar:

- Im „*Allgemeinen Wissen (AW)*“ soll dasjenige Wissen abgefragt werden, „das sich ein Durchschnittsmensch mit durchschnittlichen Bildungsmöglichkeiten selbst aneignen kann“ (Matarazzo, 1982, S. 272). „Die Wissensbereiche (...) sind bewußt sehr heterogen gehalten“ (Handbuch, S. 16), daneben jedoch auch - wohl eher unbewußt - die „Wissensarten“: So wird neben deklarativem Wissen (Hauptstadt der Türkei) auch „Prozeßwissen“ („Wie bewirkt die Hefe das Aufgehen des Kuchenteiges?“) abgefragt.

Sehr problematisch erscheinen mir zwei, *problematisch* ein weiteres Item aus AW.

- *Zu den beiden sehr problematischen:* In ihnen wird nach der Zahl der Bundesländer bzw. nach der Zahl der Einwohner Deutschlands gefragt. Aus dem Handbuch ist nicht ersichtlich, wann die Normdaten erhoben worden sind; die Quoten für die Schulabschlußgruppen sind dem Statistischen Jahrbuch 1986 (1987) entnommen, dies ist auch die jüngste ins Literaturverzeichnis aufgenommene Quelle. Ende 1991 wurde der Test veröffentlicht: Entweder wurde die Normierungsgruppe exakt zu einer Zeit befragt, als die Medien - nicht nur in unserem Land - fast täglich von 80 Millionen Deutschen und den 16 Bundesländern berich-

teten, oder die Itemkennwerte - und damit letztlich die IQ-Schätzungen - wurden mit dem Prozeß der Vereinigung Deutschlands partiell obsolet.

- *Zum problematischen Item:* „Wer wählt bei uns den Bundeskanzler?“

M.E. ist es ein handwerklicher Fehler, ein Wissenselement in einen Test aufzunehmen, von dem man weiß, daß darüber alle 4 Jahre ausführlich in Funk, Fernsehen und Zeitungen berichtet wird: Die Aufgabe ändert ihren Meßbereich im Laufe einer Legislaturperiode - bzw. bei einem eventuellen konstruktiven Mißtrauensvotum.

- *Bilderergänzen (BE):* . . . In der Aufgabengruppe verblieben ist . . . ein Item (Nr. 7, HAWIE-R; Nr. 10, HAWIE), das einen schräg über einem Wasserglas schwebenden, halb gefüllten Wasserkrug zeigt; der Proband soll bemerken, daß in dieser Position eigentlich aus dem Krug fließendes Wasser eingezeichnet sein müßten. Nun vermittelt diese Abbildung jedoch eine Art „double bind information“: Einerseits ist ganz offensichtlich ein Zustand der Schwerelosigkeit dargestellt, der Krug schwebt frei, warum also sollte das Wasser - wenn überhaupt - ausgerechnet in Richtung Glas fließen? Andererseits folgt das Wasser im Krug partiell den Gesetzen der Schwerkraft, die Wasseroberfläche verläuft nämlich parallel zur Oberfläche des Tisches, auf welchem das Glas steht. Die ökonomischste und damit „intelligenteste“ Antwort wäre unter diesen Bedingungen also ein Hinweis darauf, daß die Wasseroberfläche im Krug nicht parallel zum Tisch verlaufen dürfte - so ließe sich der Widerspruch zwischen den Einzelinformationen mit dem geringsten kognitiven Aufwand lösen...

- *Bilderordnen (BO):* In Item 3 streiten sich zwei Knaben um ein Heft mit dem Titel ATOM COMICS, der ohne tiefere Bedeutung für den Lösungsprozeß ist und gerade deswegen entfernt werden sollte, da durch ihn provozierte Verwirrungen bei dem einen oder anderen Probanden vorprogrammiert erscheinen...

- Der *Wortschatztest (WT)* erscheint deutlich verbessert. Die meisten mehrdeutigen Worte (z. B. Kurs) wurden entfernt. Allerdings verblieb „Oase“ sowohl in seiner rein denotativen als auch in seiner metaphorischen Bedeutung im Test, und wer „adäquat“ mit „fast gleich“ erklärt, bekommt recht.

Im *Rechnerischen Denken (RD)* sitzen die Autoren nach wie vor der Illusion auf - bzw. setzen sie beim Bearbeiter voraus -, die Beziehung zwischen der Zeit, die zur Erledigung einer Aufgabe benötigt wird, und der Zahl der eingesetzten Kräfte sei grundsätzlich umgekehrt proportional (Aufgabe 14): Um in einer Sekunde zu bewerkstelligen, was eine Kraft in einer Stunde erledigt, müsse man lediglich die Zahl der Kräfte verdreitausendsechshundertfachen.

- Zum *Allgemeinen Verständnis (AV)* bedarf es eines hinführenden Satzes: Im sechsbändigen „Großen Wörterbuch der Deutschen Sprache“ findet man das Wort WARUM erläutert durch „Aus welchem Grund?“, das Wort WOZU durch „Zu welchem Zweck?“ - und an den meisten Stellen des Tests, an denen eine Warum-Frage gestellt ist, wird auch mit einer Darlegung von Gründen gerechnet (z.B. Frage 11 im „Allgemeinen Wissen“:

„Warum wärmt dunkle Kleidung besser als helle?“). Nicht so bei Item 5 aus dem „Allgemeinen Verständnis“: „Warum muß man Steuern zahlen?“ Wer hier den Grund nennt, nach dem eindeutig gefragt ist, also die zweifelsohne richtige Antwort gibt: „Weil es Gesetz ist“, erhält null Punkte. Die richtige Beantwortung der gar nicht gestellten Frage, wozu Steuereinnahmen gebraucht wurden, erbringt dagegen die volle Punktzahl, zwei Punkte. Hier hat der Testautor die unter Philosophen und Handlungstheoretikern kontrovers diskutierte Frage nach Kausalität und Teleologie von Handlungen entschieden - zuungunsten des sprachlich differenzierten Testbearbeiters: Auch wenn wir umgangssprachlich Warum-Fragen stellen, um den Zweck einer Handlung zu erfahren, können und dürfen wir nicht den „bestrafen“, der in der Lage ist, die sprachlich sinnvolle Unterscheidung zu vollziehen...

Nun könnte sich der Testkonstrukteur auf die Position einer trennschärfegeleiteten Auswahl von Aufgaben und deren jeweiliger Lösung zurückziehen: Jene Antwort auf Item A ist „intelligent“, die von jenen Probanden gegeben wird, die in der komplementären Itemmenge maximale Werte erzielt haben... Was passiert aber, wenn diese ansonsten intelligente Gruppe zu dem in Item A dargestellten Sachverhalt eine Fehlvorstellung verinnerlicht hat?

- Dieser Fall ist beim ersten Item des Untertests *Gemeinsamkeitenfinden* (GF) eingetreten. Gefragt wird nach der Gemeinsamkeit von Apfelsine und Banane, und als eine von mehreren Lösungen wird „Baumobst“ anerkannt. Darf, soll, muß eine solchermaßen falsche Antwort - Bananen wachsen nun nicht auf Bäumen, sondern auf Stauden (für Skeptiker siehe Rauh, 1950, S. 254) - dann als richtig anerkannt werden, wenn die Trennschärfe (.46 bei $p = .70$; ob part-whole-korrigiert, ist dem Handbuch nicht zu entnehmen) besagt, daß die ansonsten in diesem Untertest „Guten“ dieser irrigen Auffassung sind? Intelligenztests haben nicht nur eine deskriptiv-messende Funktion, sie wirken auch präskriptiv in dem Sinne, daß sie operationale Definitionen von Intelligenz sind, also festschreiben, was intelligentes Verhalten ist...

Nach wie vor liegt dem HAWIE *Wechslers Definition von Intelligenz* zugrunde als der zusammengesetzten oder globalen „Fähigkeit des Individuums, zweckvoll zu handeln, vernünftig zu denken und sich mit seiner Umwelt wirkungsvoll auseinanderzusetzen“ (Wechsler, 1964, S. 13) - und eine Auseinandersetzung mit der Umwelt ist mir dann am wirkungsvollsten möglich, wenn ich die Einstellungen, Meinungen, Vorurteile, Fehlvorstellungen der als „gut“, als „intelligent“ Etablierten zumindest kenne, wenn auch nicht unbedingt teile. In einem gewissen Widerspruch zu diesem Konstruktionsprinzip steht jedoch die Tatsache, daß die Meßbereiche der einzelnen Untertests auch inhaltlich definiert sind. So soll mit „Gemeinsamkeitenfinden“ Aufschluß über die logische Struktur der Denkprozesse“ (Handbuch, S. 17), über allgemeines und v. a. sprachliches Abstraktionsvermögen gewonnen werden - und dieses wie-

derum ist unabhängig von den Fehlvorstellungen einer wie auch immer definierten Gruppe. Wir stehen somit vor dem klassischen Dilemma einer theoriegeleiteten Testkonstruktion: einerseits der psychologisch schlüssigen Operationalisierung dessen, was sprachliches Abstraktionsvermögen sei, und der Konfrontation dieser Überlegungen mit empirisch erhobenen Daten, andererseits der Einsicht, daß das Verhaftetsein in einer offensichtlichen Fehlvorstellung von biologischen Gegebenheiten als (Mit)Indiz für ein überdurchschnittlich ausgeprägtes sprachliches Abstraktionsvermögen herangezogen werden kann. Ich meine, die Entscheidung ist eindeutig: Es gibt Sätze, über deren Richtigkeit sich unschwer Konsens herstellen läßt; sie lassen sich nicht per Mehrheitsbeschluß von „richtig“ in „falsch“ umdefinieren. Das ist bei naturwissenschaftlichen Sachverhalten wohl ebenso unstrittig wie bei den Ergebnissen der Rechenaufgaben in diesem Test. Es sollte jedoch auch vor der Testkonstruktion klar definiert sein - das muß nicht bedeuten, daß es in der Fachöffentlichkeit unumstritten ist -, was unter dem zu messenden Konstrukt zu verstehen sei und was die im Sinne dieses Konstrukts „richtigen“ Antworten sind. Die A-posteriori-Zuordnung einer Antwort zur Gruppe der richtigen allein aufgrund statistischer Überlegungen ist - wie das Beispiel zeigt - zumindest problematisch, nur wird es uns nicht immer so deutlich vor Augen geführt wie in diesem Fall. Und ein weiteres Dilemma wird für mich an diesem Item deutlich: Das Dilemma, daß Konstrukteure von Intelligenztestaufgaben - aber beileibe nicht nur sie - stillschweigend davon ausgehen, daß ihre persönliche „Ladung“ auf der Fähigkeitsdimension, zu deren Messung sie gerade einen Test konstruieren, nicht mehr transzendierbar sei.“

9.3.3 Resümee zu Kapitel 9.3

Es wurde ein Katalog von Kriterien referiert, den das „Testkuratorium“ in der Absicht erstellt hat, den Psychologen die Bewertung von Tests zu erleichtern.

Es wurde eine Test-Rezension des HAWIE-R übernommen - als Beispiel für eine Bewertung, die sich an dem Kriterienkatalog orientiert hat.

9.4 Kontrollfragen zu Kapitel 9

- Definition von Leistungstest.
- Unterschiedliche Aspekte des globalen Leistungsbegriffes.
- Klassifikation von Leistungstests.
- Beitrag zur Diagnostik.
- Beitrag zur Intervention.
- Profilanalyse.
- Profilvergleich.

10. Kapitel

Persönlichkeitstests, Fragebogen, Persönlichkeitsinventare

Kapitel 10 bespricht eine zweite Verfahrensklasse, zu deren Verständnis „Grundkenntnisse“ vorausgesetzt werden, wie Teil II sie behandelt hat: vor allem Vertrautheit mit Testtheorie(n), aber auch mit Gesprächsführung und mit Verhaltensbeobachtung. Es geht um Persönlichkeitstests, Fragebogen, Persönlichkeitsinventare (Angleitner & Wiggins, 1986).

Die Bezeichnung „Persönlichkeitstest“ könnte suggerieren, das Verfahren erfasse die gesamte Persönlichkeit. Keineswegs erbringen „Fragebogen“ eine so umfassende Leistung, sie messen Teilbereiche einer Person: Persönlichkeitsmerkmale.

Wir fragen: Was trägt der Persönlichkeitstest zu Diagnostik und Intervention bei?

In fünf Abschnitten suchen wir nach einer Antwort:

- Abgrenzungen: Eigenart des Fragebogens (10.1),
- Bearbeitung von Fragebogen (10.2),
- Klassifikation von Fragebogen (10.3),
- Vor- und Nachteile von Fragebogen (10.4),
- Beitrag zu Diagnostik und Intervention (10.5).

Den Abschluß bilden eine Zusammenfassung des Kapitels (10.6) und eine Reihe von Kontrollfragen (10.7).

10.1 Abgrenzungen: Eigenart des Fragebogens

Persönlichkeitstest bezeichnet in diesem Kapitel eine Liste von **Feststellungen**, die nach den Regeln der (oder einer) Testtheorie formuliert wurden und gemäß vorgegebenen Alternativen zu beantworten sind. Die „Feststellungen“ ermöglichen eine formalisierte Selbstbeschreibung.

„Im Bereich der Persönlichkeitsdiagnostik haben Fragebogen die Funktion, mit Hilfe der Selbstbeurteilung Angaben über mehr oder weniger

genau umschriebene Bereiche der Persönlichkeit zu erhalten“ (Häcker; 1994 a, 256).

Kapitel 10.1 behandelt drei Themen:

- Gemeinsamkeiten mit und Unterschiede zu Leistungstests (10.1.1),
- Konstruktion von Fragebogen (10.1.2),
- Sprachregeln zur Formulierung von Fragebogen-Items (10.1.3).

10.1.1 Gemeinsamkeiten mit und Unterschiede zu Leistungstests

Der Fragebogen teilt mit dem Leistungstest bestimmte Eigenschaften, unterscheidet sich aber auch in wesentlichen Punkten von ihm.

Gemeinsamkeiten: Wie auf den Leistungstest trifft auf den Fragebogen zu:

- Er sieht *standardisierte Situationen* zur Erfassung einer Verhaltensstichprobe vor.
- Die Verhaltensstichprobe, die er zieht, gilt als *Indikator einer Personeneigenschaft*.
- Ein Proband wird charakterisiert durch Bestimmung der *relativen Position* seines Test-Scores in der Werteverteilung einer Normstichprobe.

Unterschiede: Anders als beim Leistungstest gilt beim Fragebogen:

- Die Verhaltensstichprobe besteht nicht aus ‚Realisationen‘, sondern aus ‚Deskriptionen‘ des Zielmerkmals.
- Antworten auf die Fragebogen-Items sind nicht richtig oder falsch, fallen vielmehr *in eine Schlüsselrichtung* oder *gegen eine Schlüsselrichtung* aus. - Antworten, die in Schlüsselrichtung ausfallen, tragen (zur Beschreibung und in diesem Sinne) zur Erfassung des Zielmerkmals bei. Der Testautor sieht Gründe, eine solche Zuordnung zu treffen.

Kasten 10-1 gibt Item-Beispiele aus zwei „klassischen“ Fragebogen, Kasten 10-2 soll den Unterschied zwischen Leistungs- und Persönlichkeitstest veranschaulichen.

Kasten 10-1: Fragebogen-Items: Beispiele aus bekannten Fragebogen

EPI (Eysenck Personality Inventory: Deutsche Fassung, Form A. Eggert, 1983). Angeführt werden zwei Items aus der „Neurotizismusskala“. Die Antworten lauten Ja oder Nein. (Das „X“ zeigt die Schlüsselrichtung der Auswertung an: in Richtung auf Neurotizismus.)

Skala Neurotizismus:

Item 9: „Fühlen Sie sich manchmal ohne Grund einfach miserabel?“

Ja [X] Nein []

Item 28: „Wenn Sie etwas Wichtiges getan haben, haben Sie dann oft das Gefühl, daß Sie es eigentlich hätten besser machen können?“

Ja [X] Nein []

16 **PF** (16-Persönlichkeits-Faktoren-Test. Schneewind, Schröder & Cattell, 1983). Angeführt werden zwei Items aus der Skala „Sachorientierung versus Kontaktororientierung“. Es gibt drei Antwortalternativen: (a) Nein, (b) Unsicher oder Dazwischen, (c) Ja. (Das „X“ zeigt die Schlüsselrichtung der Auswertung an: in Richtung auf Kontaktororientierung.)

Skala A: Sachorientierung versus Kontaktororientierung

Item 6: „Es wäre interessanter, ein Versicherungsagent als ein Landwirt zu sein.“

- a) Ja. [X]
- b) Dazwischen.
- c) Nein.

Item 9: „Ich wäre lieber

- a) in einem Verkaufsbüro beschäftigt,
wo ich organisieren und Leute treffen kann. [X]
- b) Dazwischen.
- c) Architekt, der in einem ruhigen Raum Pläne zeichnen kann.”

Kasten 10-2:

Unterschied zwischen Leistungstest und Fragebogen: Realisation gegenüber Deskription

1. Im **HAWIE-R** (Tewes, 1991, 62) gehört zum Subtest ‚Rechnerisches Denken‘ ein Item, das lautet: „Ein Hemd kostet 60 DM. Im Schlußverkauf wird der Preis um 15 Prozent gesenkt. Was kostet das Hemd im Schlußverkauf?“ - (Lösung: 51 DM.)
2. Ein (fiktiver) **Persönlichkeitstest** erfasse ebenfalls Rechnerisches Denken. Ein Item könnte lauten: „Ich finde es leicht, Dreisatzaufgaben zu lösen.“ - Antwort: Ja [] Nein [].

Um Rechnerisches Denken zu bekunden, muß der Proband

- im Falle 1 eine *Rechenoperation* durchführen,
- im Falle 2 dagegen *nur eine Beschreibung* abgeben, er braucht nur anzukreuzen, was er für zutreffend hält, ein Ja oder ein Nein.

10.1.2 Konstruktion von Fragebogen

Die Konstruktion von Fragebogen verläuft im wesentlichen nach den Regeln, welche die Genese eines Tests bestimmen (siehe Kap. 4, S. 31; vgl. Mummendey, 1987, 89-109; Tränkle, 1983, 238-240; 1993).

Wir zitieren - in Kasten 10-3 - vier typische Strategien, die Angleitner anführt (1991, 192-194; vgl. Angleitner, 1976; Angleitner & Wiggins, 1986; Buss & Craig, 1980; Amelang & Zielinski, 1994, 97-98).

Kasten 10-3:

Vier Strategien einer Fragebogenkonstruktion

(Angleitner, 1991, 192-194)

Genannt werden vier Strategien einer Fragebogenkonstruktion, die einander ergänzen:

1. Autoren, die ein neues Verfahren entwickeln wollen, identifizieren selber Verhaltensweisen, die das angezielte Merkmal kennzeichnen; sie formulieren selber die entsprechenden Items.
2. Eine Gruppe von Experten, wenn möglich aus verschiedenen Kulturkreisen, benennt eine Anzahl von Verhaltensweisen, die das angezielte Merkmal charakterisieren.

3. Aus dem Kreis der Personen, für die das neue Verfahren gelten soll, werden Vertreter gebeten, Verhaltensweisen zu nennen, die für das angezielte Merkmal typisch sind. Diese Nennung kann auf zwei Quellen zurückgehen: auf die unmittelbare *Beobachtung* ‚dominanten Verhaltens‘ oder auf die *Erinnerung* an Verhaltensweisen ‚dominanter Menschen‘.

„Subjects are given, for instance, the instruction to think of the three most dominant individuals they know and to write down five specific acts or behaviours that these individuals have performed that reflect or exemplify their dominance. For instance, a typical example of a behavioural act of dominance is ‚He / she sets goals for a group‘, (Angleitner, 1991, 193).

4. Die so erzeugten Itemlisten werden Experten (judges) vorgelegt, die ihrerseits abschätzen sollen, wie zutreffend ein Item das angezielte Merkmal charakterisiert. Die Eigenschaft eines Items, für ein Merkmal charakteristisch zu sein, erhält den klangvollen Titel *Prototypikalität*. Aufgabe der Experten ist es demnach, die Prototypikalität jedes Items einzuschätzen - ein Paradigma inhaltlicher Validierung (siehe Kap. 4, S. 95).

„We may conclude that after passing the prototypicality test, such lists containing highly prototypical behavioural manifestations may be regarded as possessing a much higher degree of exhaustiveness and representativeness than the lists generated by the two strategies mentioned above. Exhaustiveness, representativeness, and objectivity would be maximized if one could collect all possible behavioural manifestations that people are able to express in their languages. This, however, would be impossible“ (Angleitner, 1991, 193).

Stärke und Schwäche des Fragebogens dürften klar geworden sein. Er liefert eine nach den Regeln der (oder einer) Testtheorie formalisierte Selbstbeschreibung. Ein Fragebogen ist charakterisiert durch seinen „*Selbstaussagecharakter*“ (Mittenecker, 1982, 102; vgl. Mummendey, 1987, 17-19, 21)

10.1.3 Sprachregeln zur Konstruktion von Fragebogen

Bei der Konstruktion von Fragebogen spielt die sprachliche Fassung der Items eine bedeutsame Rolle - mehr als bei der Konstruktion von Leistungstests.

Seit Fragebogen konstruiert werden, haben Autoren auch Regeln gesammelt, an denen sich die Item-Formulierung orientieren soll (Angleitner & Wiggins, 1986; Edwards, A.L., 1953; Lennertz, 1973; Mummendey, 1987; Tränkle, 1982, 1993; Wottawa, 1980).

Doch gilt es, solche Regeln anzuwenden mit Blick auf die Zielgruppe des Fragebogens. Die sprachliche Fassung eines Fragebogens muß adressatenbezogen sein - sie ist in diesem speziellen Sinne „stichprobenabhängig“ (Mummendey, 1987, 64; Tränkle, 1993, 245). Kasten 10-4 bringt Beispiele für Sprachregeln.

Kasten 10-4:
Sprachregeln zur Formulierung von Fragebogen-Items

1. Die Formulierung der Items soll sich ausrichten „an der Alltagssprache (Umgangssprache) des durchschnittlichen Mitgliedes der Zielpopulation“ (Tränkle, 1982, 262).
2. Die Items sollen nur klare Begriffe enthalten. „Begriffe eignen sich in dem Maße, in dem sie denotativ eindeutig und konnotativ arm sind“ (Tränkle, 1993, 245).
3. Die Sätze sollen „kurz sein und nur selten mehr als zwanzig Wörter enthalten“ (Mummendey, 1987, 63).
4. Ein Item sollte nur *einen* Sachverhalt betreffen.
5. Die Items sollen Augenscheinvalidität provozieren, um den Probanden zur Mitarbeit zu motivieren.
6. Die Items sollen keine Antwort nahelegen, in der sich ‚soziale Erwünschtheit‘ ausdrückt.
7. Die Items sollen ausbalanciert gepolt sein; d.h. bei dem einen Teil der Antworten soll ein ‚Ja‘, bei dem anderen ein ‚Nein‘ in Schlüsselrichtung liegen.
8. Die Items sollen frei sein von Suggestionen.
9. Die Items sollen Häufigkeiten (z.B. oft, selten, mehrmals) nicht in Worten angeben, sondern in Zahlen oder in Beispielen. (Ausnahme: ‚Immer‘ und ‚Nie‘!)
10. Der Autor soll vermeiden im Wortgebrauch:
 - lange Wörter,
 - ungebräuchliche Wörter,
 - erst recht ungewöhnliche Fremdwörter,
 - Fachtermini.
11. Der Autor soll vermeiden im Satzbau:
 - Doppelfragen,
 - doppelte Verneinungen,
 - ungewöhnliche Satzkonstruktionen (etwa Schachtelsätze),
 - ungewöhnliche Tempora (etwa Plusquamperfekt oder Konjunktiv des Imperfekts),
 - passivische Formulierungen.
12. Die Items sollen keine ungewöhnlichen Sachverhalten oder Situationen vorgeben.

10.2 Zur Beantwortung von Fragebogen

Gefragt sei

- nach der Kompetenz des Probanden zur Selbstbeschreibung (10.2.1),
- nach seiner Bereitschaft zur Selbstbeschreibung (10.2.2),
- nach der Relation zwischen Selbstbeschreibung und Verhalten (10.2.3).

10.2.1 Kompetenz zur Selbstbeschreibung

Fragebogen leiten den Probanden an, seine ‚Selbstbilder‘ abzurufen und sich anhand der vorgegebenen Items zu beschreiben. Eine Selbstbeschreibung schließt erhebliche Schwierigkeiten ein. So kann ein Item Antworten auf unterschiedlichen Ebenen erfordern, beispielsweise auf der Verhaltens-, auf der Meinungs- oder auf der Gefühlsebene (Seitz & Rausche, 1976, 6).

Beispiel: Erfast werden soll die *Geselligkeit* eines Menschen.

- Ein Item, das sich auf die **Verhaltensebene** bezieht, könnte lauten: „*Bei gesellschaftlichen Veranstaltungen bleibe ich im Hintergrund.*“
- Ein Item, das sich auf die **Meinungsebene** bezieht, könnte lauten: „*Von gesellschaftlichen Veranstaltungen halte ich nicht viel.*“
- Ein Item, das sich auf die **Gefühlsebene** bezieht, könnte lauten: „*Bei gesellschaftlichen Veranstaltungen fühle ich mich unwohl.*“

Über welche der drei Ebenen kann der Befragte am verlässlichsten urteilen? Doch wohl über die Verhaltensebene! Denn über sein konkretes Verhalten weiß der Befragte am besten Bescheid.

Doch selbst auf der Verhaltensebene stellt eine Antwort den Probanden vor eine typische Schwierigkeit: Ein Fragebogen-Item fordert ihn dazu auf, eine Art ‚Mittelwert‘ zu bilden. Warum? Weil - um das Beispiel erneut zu zitieren - die Verhaltensweise, ‚bei sozialen Veranstaltungen im Hintergrund zu bleiben‘, in zehn Situationen zehn verschiedene Formen annehmen kann!

Beispiele für unterschiedliche Situationen: *In der ersten Situation, in der Schulklasse, zeige ich bei der Frage eines Lehrers nie auf. Ist aber die Klassensituation überhaupt eine gesellschaftliche Veranstaltung?* - *In der zweiten Situation, auf einer Klassenparty warte ich als Junge darauf ob ein Mädchen mich zu einem Tanz auffordert. Eine Party ist gewiß eine gesellschaftliche Veranstaltung!* - *In der dritten Situation, beim Einkaufen, lasse ich mich in der Schlange vor der Kasse von einer forschenden Hausfrau zurückdrängen. Erneute Frage: Handelt es sich um eine gesellschaftliche Veranstaltung? Wohl kaum! Aber es handelt sich um eine Situation, in der ‚ich mich zurückhalte‘!* - *In der vierten Situation, auf der Hochzeit meiner Kusine, sitze ich neben einer Partnerin, deren Geschwätzigkeit mich stumm macht.* - *In der fünften Situation, bei der Diskussion mit Fachkollegen, lasse ich mich überrollen von der Vielzahl der Argumente meines Kollegen A.* - *In der sechsten Situation...*

Wenn der Proband das Item beantworten soll „*Bei gesellschaftlichen Veranstaltungen bleibe ich im Hintergrund*“ muß er viele Situationen vergleichen, vermutlich mehr als zehn. Die Verhaltensvarianten kann er in seiner Antwort nicht ausdrücken - der Fragebogen sieht dies nicht vor. Der Proband muß die Verhaltensanteile gleichsam gewichten und sie in einem Ja oder Nein zusammenfassen, in diesem Sinn also eine Art ‚Mittelwert‘ bilden.

Diesen Sachverhalt spezifiziert Mummendey aus einer interaktionistischen Sichtweise: „Fragebogen geben uns Aufschluß darüber wie Personen in ganz bestimmten (Untersuchungs-)Situationen über ihr Verhalten und Erleben, ihre Einstellungen und ihre Auffassungen von sich berichten, Mit ihnen erfassen wir nicht nur als überdauernd angesehene Reaktionsweisen von Personen auf schriftlich präsentierte Situationen, sondern wir erfassen gleichfalls die subjektive Interpretation dieser im Fragebogen vorgegebenen Situationen“ (1987, 49).

Auch im günstigsten Falle - bei Fragen zum konkreten Verhalten - liegt dem Probanden das Zielmerkmal keineswegs klar und eindeutig vor Augen, so daß er gleichsam nur zugreifen könnte: Bevor er antwortet, muß er sortieren, interpretieren, eliminieren...

Wenn es schon schwierig ist, konkretes Verhalten kurz und eindeutig einzustufen, so verschärft sich die Schwierigkeit auf der Ebene von Meinung und Gefühl: Meinungs- und Gefühlsanteile einzuschätzen und richtig wiederzugeben ist erheblich schwieriger.

10.2.2 Bereitschaft zur Selbstbeschreibung: Antworttendenzen

Angenommen, der Proband **kann** sich zutreffend beschreiben: **Will** er dann auch tun, was er kann? Statt die Antworten an seinen Selbstbildern auszurichten, kann er sie orientieren an anderen Maßstäben, beispielsweise an der Tendenz,

- von sich *ein sozial erwünschtes Bild* zu bieten,
- *bestimmte Merkmale hervorzukehren* oder aber zu verschweigen,
- einem vorgegebenen *Antwortmuster* zu folgen.

Dabei wird die ‚Bereitschaft‘, im Sinne solcher Tendenzen zu antworten, ihrerseits als eine besondere ‚Personen-Eigenschaft‘ (eine Art trait) interpretiert.

Diese ‚Selbstpräsentation‘ läßt sich als eine Spielart des ‚Impression-Managements‘ beschreiben: Eine Person ist stets bestrebt, das Bild mitzugestalten, das andere Personen von ihr „haben“ (sollen). Mit diesem Streben muß sie weder Verstellungs- noch Täuschungsabsichten verknüpfen. Sie verhält sich ständig wie ein Schauspieler der sich mehr und weniger bewußt an seine Rolle anpaßt (Mummendey, 1987, 197; vgl. Graumann, 1972, 1229).

Die Forschung zu Antworttendenzen und die Literatur über diese Forschung ist umfangreich (Keil, 1973; Mittenecker, 1982, 98-105; Mummendey, 1987, 143-193). Aber wie groß „der Einfluß dieser Antworttendenzen auf den Testwert ist, läßt sich sehr schwer abschätzen, da einige Autoren postulieren, daß ein Großteil der Testwertvariation zu Lasten dieser Tendenzen gehe, während andere Autoren den Einfluß verneinen“ (Leichner, 1983, 82).

Zu unterscheiden sind die Kontrolle der Antworttendenzen

- bei der *Konstruktion* von Fragebogen (A),
- bei der *Anwendung* von Fragebogen (B).

(A)

Kontrolle von Antworttendenzen bei der Fragebogen-Konstruktion

Bei der Konstruktion versuchen Autoren auf unterschiedliche Art, diese Tendenzen zu kontrollieren. Einige Beispiele:

- Die **Tendenz**, im Sinne **sozialer Erwünschtheit** zu antworten, wird mit eigenen Skalen gemessen. Ist die Tendenz zu stark, kann der Untersucher unterschiedlich vorgehen:
 - ⇒ Er kann die Berechnung von *Korrekturwerten* vorsehen, welche die Ergebnisse auf anderen Skalen ausbalancieren.
 - ⇒ Er kann aber auch darauf *verzichten, die Fragebogenergebnisse zu verwerfen* (Edwards, A. L., 1953; Lück & Timaeus, 1969, 134-141; Mittenecker, 1982, 101; Mummendey, 1987, 159-190; Stumpf et al., 1985, 10-11).
- Die Tendenz, Merkmale zu verschweigen oder hervorzukehren, die **Dis simulations- oder Simulationstendenz**, wird mit sogenannten Offenheitskalen gemessen. (Man spricht auch von „Lügenskalen“, doch ist diese ethische Kennzeichnung unangemessen.) Ist die Tendenz zu stark, kann der Untersucher vorgehen wie eben beschrieben:
 - ⇒ Wieder kann er *Korrekturwerte* berechnen. So ist es vorgesehen beim ‚Minnesota Multiphasic Personality Inventory‘ (MMPI: Hathaway & McKinley, 1951, 1963; vgl. Dahlstrom et al., 1972).
 - ⇒ Wieder kann er *auf eine, Verwendung‘ verzichten*. So ist es vorgesehen bei der deutschen ‚Personality Research Form‘ (PRF: Stumpf et al., 1985, 82-83; vgl. auch FPI: Fahrenberg et al., 1978; Mittenecker, 1982, 98-100; Mummendey, 1987, 160).
- Die Neigung, einem vorgegebenen Antwortmuster zu folgen, die **Ja-Sage-Tendenz**, läßt sich in mindestens zwei Antwortstile aufgliedern. Es lassen sich unterscheiden:
 - ⇒ erstens eine Tendenz, bestimmten *inhaltlichen Mustern zuzustimmen* und in diesem Sinne nach sozialer Konformität zu streben,
 - ⇒ zweitens eine Tendenz, *„mechanisch“ bestimmten Antwortmustern zu folgen*, z. B. immer ein Ja oder immer ein Nein anzukreuzen (vgl. Mummendey, 1987, 146-151).

Der Effekt dieser Zustimmungstendenz läßt sich verringern, wenn die Items unterschiedlich gepolt werden: wenn die ‚Antwort in Schlüsselrichtung‘ einmal bei einem Ja, einandermal bei einem Nein liegt (Cronbach, 1946, 1950; Mittenecker, 1982, 102; Rorer, 1965).

Kasten 10-5 bringt Beispiele für die Messung dreier Antworttendenzen.

Kasten 10-5:
Skalen zur Messung von Antworttendenzen

1. Antworttendenz SOZIALE ERWÜNSCHTHEIT

SDS-E (Social Desirability Scale-Edwards: Lück & Timaeus, 1969). Die Antworten lauten „Richtig“ oder „Falsch“. (Das „X“ gibt die Schlüsselrichtung der Auswertung an: in Richtung sozialer Erwünschtheit.)

Item 1: „Ich zögere niemals, jemandem, der in Schwierigkeiten ist, zu helfen, auch wenn ich dadurch mitten in meiner Arbeit aufhören muß.“

Richtig [X] Falsch []

Item 14: „Wenn ich etwas nicht weiß, gebe ich es ohne Zögern zu.“

Richtig [X] Falsch []

Item 20: „Manchmal bin ich neidisch, wenn andere Glück haben.“

Richtig [] Falsch [X]

2. Tendenz zur VERSCHLOSSENHEIT (unangemessen: „Lügertendenz“)

FPI (Freiburger-Persönlichkeits-Inventar: Gesamtform: Fahrenberg et al., 1978). Die Antworten lauten „Stimmt“ oder „Stimmt nicht“. (Das „X“ gibt die Schlüsselrichtung der Auswertung an: in Richtung auf Offenheit.)

Skala 9: *Offenheit versus Verschlossenheit*

Item 168: „Ich bin manchmal mürrisch und schlecht aufgelegt.“

Stimmt [X] Stimmt nicht []

Item 179: „Als Kind habe ich ab und zu mal genascht.“

Stimmt [X] Stimmt nicht []

Item 190: „Ich bin hin und wieder ein wenig schadenfroh.“

Stimmt [X] Stimmt nicht []

3. JA-SAGE-TENDENZ

FRAGEBOGEN der ZUSTIMMUNGSTENDENZ (Mummendey, 1987, 148-151). Das Verfahren besteht aus gängigen Sprichwörtern: *Gemessen wird das Streben nach sozialer Konformität*. Nicht erfaßt wird die Tendenz, ‚mechanisch‘ einem Antwortmuster im Fragebogen zu folgen, z.B. immer ein Ja oder immer ein Nein anzukreuzen. (Das „X“ gibt die Schlüsselrichtung der Auswertung an: in Richtung der Zustimmungstendenz.)

Item 1: „Alte Bäume soll man nicht verpflanzen.“ Ja [X] Nein []

Item 22: „Lügen haben kurze Beine.“ Ja [X] Nein []

Item 46: „Wes 'Brot ich eß', des 'Lied ich sing'.“ Ja [X] Nein []

(B)

Kontrolle von Antworttendenzen bei der Fragebogen-Anwendung

Bei der Anwendung eines Fragebogens kann der Diagnostiker ebenfalls auf unterschiedliche Weise versuchen, Antworttendenzen zu beachten. Auch dafür einige Beispiele:

- Die **Instruktion** bietet eine Möglichkeit, Tendenzen zu kontrollieren (vgl. Mittenecker, 1982, 99-100): Der Untersucher kann
 - ⇒ seine Bereitschaft signalisieren, das Ergebnis mit dem Probanden zu besprechen und ihm Unklarheiten zu erläutern;
 - ⇒ den Probanden auffordern, im eigenen Interesse aufrichtig zu antworten;

- ⇒ dem Probanden erklären, er irre sich, wenn er annehme, daß er sofort erkenne, welche Antwortalternative für ihn günstiger sei.
- Eine andere Kontrollmöglichkeit heißt **Einschätzung der diagnostischen Situation**. Vielleicht liegt darin sogar die größte Chance, Verzerrungstendenzen zu prüfen. Gemeint ist die Abschätzung der Motive, die den Probanden zum Diagnostiker geführt haben:
 - ⇒ Wenn ein Proband ‚von sich aus‘ einen Psychologen um Untersuchung und Beratung bittet, ist eher zu vermuten, daß er - auf die Gefahr einer Verfälschung aufmerksam gemacht - daran mitarbeitet, Informationsverzerrungen zu vermeiden.
 - ⇒ Wenn ein Proband dagegen nicht ‚von sich aus‘ kommt, sondern ‚ge-nötigt‘ ist, an einer psychologischen Untersuchung teilzunehmen, dann wächst die Gefahr der Informationsverzerrung. Gibt es eine Abhilfe? Der Diagnostiker muß Fragebogenergebnisse - *noch sorgfältiger als in anderen Fällen* - an anderen Verfahren ‚validieren‘.
- Die Auswertung von **Kontrollskalen**, die ein Fragebogen enthält, gibt eine weitere Hilfe, den Einfluß von Verzerrungstendenzen abzuschätzen. So enthält
 - ⇒ das (E-P-I) eine ‚Lügenskala‘,
 - ⇒ das FPI eine ‚Offenheitsskala‘,
 - ⇒ das MMPI vier Kontrollskalen (1. ?-Wert: Unbeantwortete Items; 2. L-Wert: Lügenskala; 3. F-Wert: Frequency Scale; 4. K-Wert: Korrekturskala).
 Die Werte solcher Skalen bei einer Interpretation der anderen Skalen mit-zuberücksichtigen kann zu einer ‚Verifikation oder Validierung im Indi-vidualfall‘ beitragen.

10.2.3 Relation von Selbstbeschreibung und Verhalten

Der Selbstaussagecharakter bringt es mit sich, daß ein Fragebogen eher Auskunft gibt über Bereitschaften, Vorstellungen, Einstellungen zum Verhalten als über tatsächliches Verhalten.

Die Beziehung zwischen Selbstbeschreibung und Handeln ist sehr differenziert zu betrachten. Verschiedene Autoren kommen „auf Grund systematischer Analysen vorliegender empirischer Studien zu dem Ergebnis..., daß die häufig unterstellte einfache Beziehung zwischen Einstellung und Handeln in keiner Weise durch das empirische Material bestätigt wird, daß im Gegenteil generell nur von einer schwachen Beziehung zwischen ihnen gesprochen werden kann“ (Meinefeld, 1983, 94; vgl. Six, 1975).

Ob zwischen Einstellung und Handeln eine Beziehung besteht, hängt ab von den Methoden, mit denen diese Beziehung untersucht wird (Meinefeld, 1983, 95):

- Selbstberichte über Einstellung und Handeln überschätzen den Zusammenhang.
- Methoden, die nur einzelne Handlungsaspekte beachten oder die vom Probanden ein Abweichen von der Handlungsroutine fordern, unterschätzen die Stärke des Zusammenhanges.

Folgerung: Fragebogenergebnisse gelten als Aussagen über Vorstellungen, Wünsche, Einstellungen zu Verhalten. *Aus Fragebogenergebnissen allein sollen keine Aussagen über tatsächliches Verhalten abgeleitet werden.*

10.3 Klassifikation von Fragebogen

Fragebogen werden unter verschiedenen Perspektiven klassifiziert. Wir folgen der Einteilung, die Brickenkamp in seinem Handbuch vorgibt (1975, 14-15); er unterscheidet

- Persönlichkeitsstrukturtests,
- Einstellungs- und Interessentests sowie
- Klinische Tests.

Persönlichkeitsstrukturtests

Der Begriff ‚Struktur‘ besagt hier, daß ein Fragebogen mehrere Persönlichkeitsmerkmale erfaßt und für diese Merkmale eine ‚Ordnung‘ annimmt - ihnen in diesem Sinne eine Struktur zuerkennt. Hinzu kommt, daß die Merkmalsausprägungen im Bereich der ‚normalen‘ Persönlichkeit auftreten.

Zwei Beispiele:

1. **Gießen-Test** (GT: Beckmann, Brähler & Richter, 1990). Der Fragebogen ist faktorenanalytisch konstruiert und erfaßt sechs Merkmale, die bipolar angeordnet sind: (1) Positive vs Negative Soziale Resonanz, (2) Dominanz vs Gefügigkeit, (3) Unterkontrolliertheit vs Zwanghaftigkeit, (4) Hypomanie vs Depressivität, (5) Durchlässigkeit vs Retentivität, (6) Soziale Potenz vs Soziale Impotenz.
2. **NEO-Fünf-Faktoren Inventar** (NEO-FFI: Borkenau & Ostendorf, 1993). Der Fragebogen ist faktorenanalytisch konstruiert und erfaßt fünf Merkmale: (1) Neurotizismus, (2) Extraversion, (3) Offenheit für Erfahrung, (4) Verträglichkeit und (5) Gewissenhaftigkeit.

Einstellungs- und Interessenfragebogen

Einstellungen und Interessen haben gemeinsam, daß sie das menschliche Orientierungsverhalten vorstrukturieren und so die Kommunikation mit der

Umwelt erleichtern. Jede Vorstrukturierung ist geprägt von der individuellen Lerngeschichte in ihren kognitiven, affektiven, konativen Komponenten.

Zu trennen sind die beiden Konstrukte eher nach Schwerpunkten:

- *Einstellungen* bezeichnen eher Meinungen, Anschauungen, Überzeugungen zu Sachverhalten,
- *Interessen* stehen Verhaltens- und Handlungstendenzen näher.

Zwei Beispiele:

1. **Berufs-Interessen-Test II** (BIT 11: Irle & Allehoff, 1984). Der Fragebogen erfaßt neun berufsorientierte Interessensrichtungen: (1) Technisches Handwerk, (2) Gestaltendes Handwerk, (3) Technische und naturwissenschaftliche Berufe, (4) Ernährungshandwerk, (5) Land- und fortwirtschaftliche Berufe, (6) Kaufmännische Berufe, (7) Verwaltende Berufe, (8) Literarische und geisteswissenschaftliche Berufe, (9) Sozialpflege und Erziehung. - Die Interessensschwerpunkte sollen erkannt werden aus der Wahl, die ein Proband trifft zwischen unterschiedlichen berufsbezogenen Tätigkeiten.
2. **Differentieller Interessen-Test** (DIT: Todt, 1967). Der Fragebogen ermittelt Interessensbereiche aus Beruf und Freizeit: (1) Sozialpflege und Erziehung, (2) Politik und Wirtschaft, (3) Verwaltung und Wirtschaft, (4) Unterhaltung, (5) Technik und Naturwissenschaft, (6) Biologie, (7) Mathematik, (8) Musik, (9) Kunst, (10) Literatur und Sprache, (11) Sport. - Die Interessensschwerpunkte werden erschlossen aus den Vorlieben für die vier Bereiche (a) Berufe, (b) Bücher, (c) Zeitschriften und (d) Tätigkeiten.

Klinische Fragebogen

„Tests, die in dieser Gruppe aufgeführt werden, sollen Anhaltspunkte geben für eine klinische Diagnosestellung, sollen psychopathologische Erscheinungen erfassen und Hilfen für eine differential-diagnostische Abklärung anbieten. Sie können - für sich genommen - aber keine psychiatrische Diagnose ersetzen“ (Brickenkamp, 1975, 15).

Zwei Beispiele:

1. **Biographisches Inventar zur Diagnose von Verhaltensstörungen** (BIV Jäger, Lischer, Münster & Ritz, 1976). Der Fragebogen soll acht sensitive Determinanten für gestörtes Verhalten erfassen: (1) Familiäre Situation, (2) Ichstärke, (3) Soziale Lage, (4) Erziehungsverhalten, (5) Neurotizismus, (6) Soziale Aktivitäten, (7) Psychophysische Konstitution, (8) Extraversion.
2. **Veränderungsfragebogen des Erlebens und Verhaltens** (VEV: Zielke & Kopf-Mehnert, 1978). Der Fragebogen erfaßt einen bipolar angeordneten Faktor: ‚Spannung, Unsicherheit, Pessimismus‘ versus ‚Entspannung, Gelassenheit, Optimismus‘. An diesem Faktor soll das Ergebnis einer Verhaltens- und Erlebensänderung nach einer Gesprächspsychotherapie ablesbar werden.

HINWEIS: Verwiesen sei auf eine **Sprachregelung**, welche die *Transparenz von Fragebogen*, damit aber auch ihre Klassifikation betrifft. Nach einem Vorschlag von Cattell gilt:

- **Objektiv** ist ein Fragebogen, wenn er so angelegt ist, daß der Proband ihn nicht durchschaut, ihn im Idealfalle also nicht verfälschen kann, Cattell hält das ‚Sixteen Personality Factor Questionnaire‘ (16 PF) für ein Paradigma objektiver Persönlichkeitstests (vgl. Cattell, 1958; Cattell, Eber & Tatsuoka, 1970; Schneewind, Schröder & Cattell, 1983).
- **Subjektiv** ist ein Fragebogen, wenn der Befragte ihn durchschauen und insofern auch verfälschen kann. Als Beispiel ließe sich jeder der ‚üblichen‘ Fragebogen anführen.

Objektiv bedeutet hier demnach etwas anderes als in der klassischen Testtheorie, nach der ein Verfahren dann als objektiv gilt, wenn Erhebung, Auswertung und Interpretation ‚standardisiert‘ und in diesem Sinne ‚unabhängig vom Untersucher‘ sind (siehe Kap. 4, S. 66).

10.4 Vorzüge und Nachteile von Fragebogen

In einer Übersicht seien Chancen und Grenzen aufgelistet, die mit dem Fragebogen gegeben sind:

- Nachteile, Grenzen, Probleme (10.4.1),
- Chancen, Vorteile, Möglichkeiten (10.4.2).

10.4.1 Nachteile, Grenzen, Probleme

Die Probleme seien zuerst mit einem Stichwort genannt, dann in unterschiedlicher Ausführlichkeit kommentiert.

Vereinfachtes Reiz-Reaktions-Schema: Konzipiert ist der Fragebogen unter der Annahme, ein Fragebogen-Item wirke wie ein Reiz, auf den ein Proband reagiere, indem er - ohne Zwischenschritte einer Interpretation - das angesprochene Merkmal abrufe. Inzwischen weiß die Psychologie, daß dieses Reiz-Reaktions-Modell - weil zu einfach - auf verbale Prozesse nicht anwendbar ist. Das Item als Reiz wird interpretiert, bevor ihm als Reaktion eine Antwort folgt.

Beispiel: *Schlagend läßt sich die individuelle Interpretation demonstrieren an Adverbien wie ‚oft‘ oder ‚selten‘: „Ich habe oft Kopfschmerzen.“ - „Ich nehme selten an Parties teil.“ Was der Einzelne meint, wenn er ‚oft‘ oder ‚selten‘ ankreuzt, bleibt unklar*

„Nach dem Prinzip der ‚maximalen Übelminimierung‘ (Wottawa, 1980, 209) werden bei der formalen Standardisierung (Identität der Fragen auf verbaler

Ebene) die interindividuellen sprachlichen Unterschiede vernachlässigt. Ziel bleibt auch hier die Bedeutungsäquivalenz, nur liegt diesem Vorgehen die Annahme zugrunde, daß diese durch identische sprachliche Formulierungen zu erreichen sei“ (Tränkle, 1982, 262).

Unterschiedliche Antwortmuster gelten als gleich: Erhalten zwei Probanden den gleichen Test-Score, dann gilt das Merkmal bei beiden als gleich ausgeprägt. Die Frage stellt sich aber, ob gleiche Test-Scores dieselbe Merkmalsausprägung „abbilden“, wenn sie auf Antwortsequenzen zurückgehen, deren Inhalte erheblich divergieren.

Beispiel: Aggressivität hat unterschiedliche Facetten; sie läßt sich beispielsweise aufgliedern in behaviorale und verbale Aggression. Jede der beiden Spielarten läßt sich erneut vielfältig aufschlüsseln. So kann behaviorale Aggressivität sich manifestieren in Gewalt gegen Sachen oder aber (eine Steigerung) in Gewalt gegen Personen. Verbale Aggression kann sich manifestieren in offenen Beschimpfungen, aber auch in ‚sanften‘ ironischen Bemerkungen.

Nun mögen zwei Itemscores gleich hoch ausfallen, etwa zwölfmal ein ‚Ja‘ für Aggressivität repräsentieren. Dann kann in dem einen Falle der Score zurückgehen auf acht Items, die behaviorale Aggression, und auf vier Items, die verbale Aggression anzeigen. In dem anderen Falle kann er beruhen auf acht Items, die verbale Aggression, und vier Items, die behaviorale Aggression anzeigen. - Repräsentieren die beiden Scores dennoch die gleiche Ausprägung von Aggressivität?

Mehrwertigkeit der Items und des Zielmerkmals: Es gibt kaum ein Item, das sich nur einem einzigen Merkmal zuordnen ließe. Meist enthält ein Item Inhalte oder inhaltliche Anteile, die Beziehungen erkennen lassen zu verschiedenen Merkmalen.

Beispiel 1:

Mehrwertigkeit eines Items: Im E-P-I (Eysenck Personality Inventory: Form A) lautet Item 28: „Wenn Sie etwas Wichtiges getan haben, haben Sie dann oft das Gefühl, daß Sie es eigentlich hätten besser machen können?“ Ein Ja gilt als Indikator für Neurotizismus. Könnte das Ja nicht ebenso auf andere - verwandte - Merkmale hinweisen, etwa auf Depressivität, auf Selbstunsicherheit, auf Perfektionismus, auf Gefügigkeit: „Eigenschaften“, die ihrerseits alle auch „selbständige“ Merkmale repräsentieren könnten?

Beispiel 2:

Mehrwertigkeit eines Zielmerkmals: Erfasst werde das Zielmerkmal ‚Extraversion/Introversion‘, eine Persönlichkeitseigenschaft, welche sich auf die Dimension Ich-Umwelt bezieht“ (Dorsch, 1994, 228; vgl. Eggert, 1983; Eysenck, 1976a; Schneewind, Schröder & Cattell, 1983).

Kein Eingeweihter zweifelt, daß es dieses Zielmerkmal „gibt“. Aber was besagt dies für die Absicht, den Bedeutungshof des Konzeptes eindeutig einzugrenzen? Das Merkmal „Extraversion/Introversion“ impliziert Aussagen über:

- *zerebrale Erregungs- und Hemmungspotentiale,*
- *Schnelligkeit des Lernens (Konditionierbarkeit)),*
- *Leichtigkeit der Aufnahme von Kontakten,*
- *sozial-verbale Gewandtheit (Witz, Schlagfertigkeit),*
- *Begeisterungsfähigkeit,*
- *Pragmatismus,*
- *Spontaneität*
- *usw.*

Ist ein Test-Autor in der Lage, die Bedeutungssegmente eines solchen Zielmerkmals anteilsgerecht in den Items ‚abzubilden‘? Wenn nicht, wie ist dann der Bedeutungshof zu bestimmen?

An dieser Stelle sei eingeräumt, daß sich eine solche Frage zwar stellen läßt angesichts der Konstruktion von Fragebogen, daß sie aber nur die generelle Schwierigkeit widerspiegelt, mit der ein Psychologe konfrontiert wird, wenn er ein Merkmal klar ‚definieren‘ will.

Unzureichende Kompetenz des Befragten: Worüber kann der Befragte in einem Fragebogen Auskunft geben? (Lassen wir die Verfälschungstendenzen sogar unberücksichtigt.)

Auskunft geben kann ein Proband darüber, was er von der ‚kognitiven Repräsentanz über sich‘ abrufen kann. Dieser Satz spricht mindestens zwei Sachverhalte an:

1. das Wissen, das einer Person über sich zur Verfügung steht;
2. die sprachliche Fähigkeit, das verfügbare Wissen auch auszudrücken.

Zu 1.: Das Wissen über die eigene Person ist begrenzt. Darum kann ein Fragebogen auch nur ein begrenztes Wissen über ein Merkmal abrufen. In der ‚kognitiven Repräsentanz über die eigene Person‘ mögen bewußte und halb bewußte Anteile von Verhaltenstendenzen ‚abgebildet‘ sein - selten jedoch die Gesamtdynamik, in die ein Merkmal eingebettet ist.

Beispiel: Die ‚soziale Potenz‘, ermittelt im Gießen-Test, bezieht sich auf die Vielfalt sozialer Bezüge, auch auf den heterosexuellen Kontakt (Beckmann et al., 1990). Sollten aber wirklich die Antworten auf sechs Fragen die Gesamtdynamik sexuellen Verhaltens widerspiegeln?

Zu 2.: Wie sorgfältig immer die Sprachgestalt eines Fragebogens abgestimmt wird auf die angezielte Population - die Probanden unterscheiden sich erheblich in der Kompetenz, ihr ‚Wissen über die eigene Person‘ in Einklang zu bringen mit dem sprachlichen Itemgehalt. Diese Frage führt zurück zu dem Problem, mit dem der kritische Überblick einge-

leitet wurde: Verbale Reize werden *vielfältig interpretiert*, bevor ihnen als Reaktion eine Antwort folgt.

Resümee: Ein Untersucher sollte nicht erwarten, in einem Fragebogen gleichsam die ‚Wahrheit‘ über einen Probanden zu erfahren. Ebenso falsch wäre eine Generalisierung im gegenläufigen Sinne, die besagte, daß ein Fragebogen für die Diagnostik nichts Nennenswertes leisten könne.

Welche Chancen der Fragebogen bietet, soll der nächste Abschnitt aufzeigen.

10.4.2 Chancen, Vorteile, Möglichkeiten

Dem Katalog von Problemen, die der Fragebogen aufwirft, sei eine Liste von Chancen gegenübergestellt.

Psychometrische Konstruktion: Zu den Vorteilen zählen wir die psychometrische Konstruktion des Fragebogens. Aus einer umfassenden Voruntersuchung liefert diese Prozedur Kennwerte (sollte sie jedenfalls liefern), die nachvollziehbar machen, worauf die Fragebogensaussagen beruhen. Insofern erleichtern sie ein Urteil über das Instrument. Kritische Urteile über Fragebogen sind nur möglich aufgrund der Analyse eben solcher Kennwerte. Soviel konkrete Vorinformation bietet kein anderes diagnostisches Verfahren - angenommen Tests im Sinne von Leistungstests.

Distanzierung vom persönlichen Eindruck: Der Fragebogen erlaubt es dem Untersucher, sich von der Untersuchungssituation, von der untersuchten Person zu distanzieren. Seine Wünsche, Ängste, Erwartungen, Stereotypen als deformierende Determinanten seines Verhaltens werden in ihrem Einfluß eingeschränkt.

Erleichterung der diagnostischen Untersuchung: Die einfache Anwendbarkeit, die rasche Auswertung, die gemeinsame Benennung, der geringe Zeit- und Materialaufwand sind Qualitäten, die den Fragebogen zu einem handlichen Instrument in der diagnostischen Situation machen.

Quantifizierung des Merkmals: Die Abbildung von Merkmalen im Zahlenrelativ erleichtert vielfältige Prüfungen, beispielsweise Vergleiche der Auswertungen, Vergleiche der verschiedenen Aspekte von Zuverlässigkeit, Vergleiche individueller und gruppenorientierter Scores.

Abbildung der Intra-Merkmalstruktur: Die interne Struktur eines Merkmals, abgebildet in den Items und ihren Scores, wird im Fragebogen überprüfbar, soweit sie sich in den Zahlenscores abbildet. Es geht um Interkorrelation, Konsistenz, Homogenität der Items. Auf diesem Wege wird deutlich, in welchen Anteilen die Feststellungen ein Merkmal erfassen und wieweit sie es in relativ selbständige Teile abgrenzen.

Abbildung der Inter-Merkmalstruktur: Ein Fragebogen, der mit anderen Verfahren verglichen wird, gibt eine doppelte Information, erstens, in welchen weiteren Verhaltenskontext ein Merkmal gehört, zweitens aber auch, in welchem Maße es sich von gleichen Merkmalen abhebt, die durch andere Verfahren gemessen werden (Multitrait-Multimethod-Validierung, S. 108).

Normbezug: In dem Vergleich des einzelnen Probanden mit einer Referenzgruppe liegt ein gewichtiges Problem, das wir bei Kritik der Klassischen Testtheorie erwähnt haben, das Problem der Populationsabhängigkeit. Hier sei jedoch auch auf die Chance verwiesen, die in diesem Bezugswert steckt. Es gibt Aussagen, die nur im Vergleich sinnvoll werden. Der Normbezug des Fragebogens ermöglicht solche sinnvollen Vergleiche. Insofern macht der Fragebogen das Individuum beschreibbar durch Vergleich mit anderen.

Sprachregelung zwischen Psychologen: Ein Fragebogen wirkt nicht nur deskriptiv, ein Vorteil liegt auch darin, daß verschiedene Untersucher für gleiche Verhaltensweisen gleiche Benennungen verwenden. Daß mit einer solchen Sprachregelung die verwendeten Konzepte schon endgültig geklärt und im theoretischen „wohldefiniert“ seien, braucht niemand anzunehmen. Es werden aber Bezugspunkte genannt, von denen her gleiche Merkmale unter gleichen Perspektiven benannt und erfaßt werden.

10.5 Beitrag zu Diagnostik und Intervention

Der Katalog der Grenzen und der Chancen, die mit dem Fragebogen gegeben sind, sollte herausarbeiten, unter welchen Aspekten sich der Fragebogen als nützlich erweist. Seine Schwächen raten davon ab, ihn isoliert einzusetzen. Seine Stärken verbieten es, auf ihn zu verzichten.

„Als Fazit kann festgehalten werden: Auf Selbsturteile grundsätzlich zu verzichten bedeutet, Erkenntnismöglichkeiten zu vergeben. Doch dürfte es wohl weniger ihr alleiniger Einsatz als vielmehr ihre Kombination und mehr noch ihr Abgleich mit anderen Quellen und Verfahren sein, der Selbsturteile diagnostisch ergiebig macht“ (Esser 1995, 654).

Eingesetzt zusammen mit Leistungstests, Verhaltensbeobachtung, Exploration und projektiven Verfahren, ermöglicht er die Anwendung des Prinzips einer konvergenten und diskriminanten Validierung (siehe Kap. 4, S. 110): Wenn ein Untersucher dieselben Merkmale mit verschiedenen Verfahrensklassen identifiziert, kann er den Bedeutungshof von Merkmalen eindeutiger verifizieren (Konvergenz) und die Merkmale schärfer gegeneinander abgrenzen (Diskriminanz).

Was leisten Fragebogen für Diagnostik, was für Intervention?

Die Stichworte können ähnlich lauten wie beim Leistungstest, sie wurden zum Teil schon kommentiert.

Für die **Diagnostik** ermöglicht der Fragebogen

- eine kontrollierte Merkmalerfassung,
- die Einbettung eines Merkmals in eine Theorie,
- eine Vielzahl von Vergleichen,
- einen Beitrag zur psychologischen Sprachregelung.

Für die **Intervention** ermöglicht der Fragebogen vor allem zwei Beiträge:

- eine Identifikation des Interventionsbedarfs und
- eine Bilanzierung des Interventionserfolgs.

Zur Veranschaulichung seien in Kasten 10-6 zwei **Beispiele** beschreiben.

Kasten 10-7:

Zwei Beispiele für die Anwendung des Fragebogens bei diagnostischen und interventiven Maßnahmen

1. Beispiel: Erfassung und Behandlung partnerschaftlicher Störungen

Quelle: „Ehe- und Partnerschaftsstörungen“ von Scholz (1987, 70-76)

Instruktion:

- „Nachfolgend finden Sie eine Anzahl von Aussagen, die partnerschaftliches Zusammenleben betreffen (z.B. Kinder, Finanzen, Haushalt).
- Jede Aussage wird durch den Satz „Meiner Meinung nach treten in unserer Ehe Schwierigkeiten im Zusammenhang damit auf, daß ...“ eingeleitet. Unter dem Wort „Schwierigkeit“ verstehen wir - auch gelegentlich auftretende - Meinungsverschiedenheiten, unterschiedliche Ansichten, Disharmonien, Konfliktpunkte bei den Partnern.
- Entscheiden Sie bei jeder einzelnen Aussage, ob Sie für Ihre Partnerschaft zutrifft = ja, oder nicht zutrifft = nein und ziehen Sie einen Kreis um die für Sie zutreffende Antwort.
- Wahrscheinlich werden einige Aussagen und Antworten auf Ihre Situation nicht vollständig zutreffen. Geben Sie dann diejenige Antwort an, die Ihre Ehe am besten beschreibt.
- Überprüfen Sie bitte zum Schluß, ob jede Frage beantwortet wurde. Nur wenn Sie *jede* Frage beantworten, sind Ihre Angaben wissenschaftlich verwertbar.“

Einige Beispiel-Items (Ziffer 14-17):

„Meiner Meinung nach treten in unserer Ehe Schwierigkeiten im Zusammenhang damit auf, daß

- | | |
|--|---------|
| 14.. wir oft voneinander nicht wissen, wie weit die gegenseitige Zuneigung geht. | ja nein |
| 15.. einer dem anderen kaum Einblick in seine Gefühle erlaubt. | ja nein |
| 16.. wir nicht immer ehrlich zueinander sind. | ja nein |
| 17.. wir uns nicht immer vertrauen können. | ja nein |

Sonstige Probleme des Vertrauens, der Offenheit: ...“

Resultat: Der Fragebogen kann das „Störungsausmaß der Ehe unter quantitativem Aspekt“ angeben (1987, 76).

2. Beispiel: Einübung von Arbeits- und Sozialverhalten

Quelle: „Training mit Jugendlichen“ von Petermann, F. & U. (1987)

Die **Ziele des Trainings** waren vielschichtig (1987, 29-36): Selbst- und Fremdwahrnehmung sollen geschärft, Ausdauer und Kontrolle „bei willentlich gelenkten Handlungen“ gestärkt (1987, 30), der Ausdruck des eigenen Körpers und der damit verbundenen Gefühle genauer wahrgenommen, Selbstsicherheit und Selbstbild erhöht, das Einfühlungsvermögen vertieft, Kritik als Lob oder Tadel angemessen bewertet, schließlich, als höchster Anspruch, die Einzelziele miteinander integriert werden.

Zur **diagnostischen Erfassung** des Interventionsbedarfs wurden unterschiedliche Instrumente herangezogen (1987,45-66): diagnostische Gespräche mit den Jugendlichen, Verhaltensbeobachtung, Aktenanalyse, Fragebogen,

Die **Fragebogen** sollten erfassen:

1. Erfahrungen von Selbstwirksamkeit und Hilflosigkeit,
2. Leistungsmotivation,
3. Berufsinteressen und Arbeitsverhalten,
4. Sozialverhalten und verschiedene Persönlichkeitseigenschaften.

zu 1.: *Selbstwirksamkeit und Hilflosigkeit* wurden ermittelt

- mit der Hilflosigkeits- und Selbstwirksamkeitsskala von Schwarzer (1981).

zu 2.: *Leistungsmotivation* wurde erfaßt

- mit dem Anstrengungsvermeidungstest (AVT) von Rollett und Bartram (1977) und
- mit dem Leistungsmotivationstest (LMT) von Hermans, Petermann und Zielinski (1978).

zu 3.: *Berufsinteressen und Arbeitsverhalten* wurden erfaßt

- mit dem Geist-Bilder-Interessen-Inventar (GBII) von Stauffer und Trottmann-Geschwend (1980) sowie
- mit dem Arbeitsverhaltensinventar (AVI) von Thiel, Keller und Binder (1979).

zu 4.: *Sozialverhalten* und verschiedene *Persönlichkeitseigenschaften* wurden erfaßt

- mit dem Angstfragebogen für Schüler (AFS) von Wiczerkowski et al. (1974) und
- mit der Hamburger Neurotizismus- und Extraversionsskala (HANES) von Bugge und Baumgärtel (1972).

Beim **Training** wurden verschiedene Verhaltensweisen eingeübt. Genannt sei nur *eine* Facette: die Einübung von „*Eigenverantwortung*“ (1987, 92-100). Der Jugendliche muß dabei

- aus einem ‚Tagebuch‘ über Aufgaben berichten, die mit Eigenverantwortung zu tun haben;
- drei konkrete Beispiele für Eigenverantwortung aus seinem Alltag schildern;
- eines dieser konkreten Beispiel in einem Rollenspiel realisieren und anhand eines Tonbandes rekapitulieren und kontrollieren.

Zur Effektkontrolle müßten naturgemäß dieselben Verfahren eingesetzt werden wie zur Diagnose. Doch verzichteten Petermann und Petermann auf Fragebogen; sie zogen nur „Beobachtungsverfahren“ heran: Expertenurteile über das Verhalten der Jugendlichen im Alltag und Trainerurteile über die Mitarbeit während der Intervention (1987, 171-177).

10.6 Zusammenfassung zu Kapitel 10

Persönlichkeitstest bezeichnet in diesem Kapitel eine Liste von Feststellungen, die nach den Regeln der (oder einer) Testtheorie formuliert wurden und gemäß

vorgegebenen Alternativen zu beantworten sind. Die „Feststellungen“ ermöglichen eine formalisierte Selbstbeschreibung.

Andere Bezeichnungen lauten Persönlichkeitstest oder Persönlichkeitsinventar. Wie bei Leistungstests wird der Test-Score verglichen mit Kennwerten, die an einer Normstichprobe ermittelt wurden.

Bei der Selbstbeschreibung kann der Proband sogenannten Antworttendenzen folgen, etwa der Tendenz,

- von sich ein Bild zu bieten, das vor allem sozial erwünschte Züge enthält,
- bestimmte Merkmale hervorzuheben oder aber zu verschweigen (Simulations- oder Dissimulationstendenz),
- einem vorgegebenen Antwortmuster zu folgen (Ja-Sage-Tendenz).

Der Selbstaussagecharakter bringt es mit sich, daß ein Fragebogen eher Auskunft gibt über Bereitschaften, Vorstellungen, Einstellungen zum Verhalten als über tatsächliches Verhalten.

In Situationen, wo Selbsterschließung und in diesem Sinne Selbstbeschreibung vom Probanden als sinnvoll erlebt werden, bieten Fragebogen eine Chance der Wahrheitsfindung.

In diagnostischen Situationen, in denen der Proband nach günstiger Selbstdarstellung streben darf (etwa bei forensischer Begutachtung oder bei Veranstaltungen zur Personalauswahl), liefert der Fragebogen Ergebnisse, die man mit Vorbehalten einstufen muß. Den Grad der „Wahrhaftigkeit“ abzuschätzen dürfte nur möglich sein, wenn der Fragebogen im Verbund mit anderen Verfahren verwandt wird.

10.7 Kontrollfragen zu Kapitel 10

Charakteristika von Persönlichkeitstests.

Gemeinsamkeiten mit Leistungstests.

Unterschiede zu Leistungstests.

Antworttendenzen.

Kontrolle von Antworttendenzen.

Klassifikation von Persönlichkeitstests.

Beitrag zur Diagnostik.

Beitrag zur Intervention.

11. Kapitel

Persönlichkeitsentfaltungsverfahren oder projektive Verfahren

Kapitel 11 behandelt eine dritte Klasse von Verfahren, zu deren Verständnis das Basiswissen aus Teil II beitragen soll - hier vorrangig die Vertrautheit mit Gesprächsführung und Verhaltensbeobachtung, aber auch die Kenntnis der Testtheorie(n).

Persönlichkeits-Entfaltungsverfahren oder projektive Verfahren sind umstritten. Den schwierigen Stoff gliedern wir in fünf große Abschnitte:

- Abgrenzung des Konzeptes der Projektion (11.1),
- Klassifikation projektiver Verfahren (11.2),
- Probleme projektiver Verfahren (11.3),
- Beitrag projektiver Verfahren zur Diagnostik (11.4),
- Darstellung von drei Klassen projektiver Verfahren (11.5).

Das Kapitel schließt mit einer Zusammenfassung (11.6) und einer Reihe von Kontrollfragen (11.7).

11.1 Abgrenzung des Konzeptes der Projektion

Auch heute noch werden in der diagnostischen Situation sogenannte Persönlichkeits-Entfaltungsverfahren oder projektive Verfahren eingesetzt (Schober, 1977), obwohl jedem Kundigen ihre Problematik vertraut ist. Ihre Objektivitäts-, ihre Reliabilitäts- und Validitätskoeffizienten liegen niedrig. „Der aus diesen Gründen zu beobachtende Verzicht auf solche Verfahren erscheint jedoch nicht gerechtfertigt, da projektive Verfahren Informationen liefern, die durch strukturierte Tests nicht zu gewinnen sind“ (Leichner, 1983, 83).

HINWEIS: Um der Kürze willen bleiben wir im weiteren Verlauf des Kapitels 11 bei der **Benennung ‚projektive Verfahren‘** und verwenden nicht mehr den alternativen Titel ‚Persönlichkeits-Entfaltungsverfahren‘.

Frage einer Definition

Was sind projektive Verfahren? Der Begriff bleibt theoretisch unscharf (Hörmann, 1982, 177-178). ‚Projektiv‘ läßt sich unterschiedlich interpretieren.

Zum einen kann das Adjektiv einen **Namen** abgeben und eine Klasse von Verfahren benennen, ohne einen gemeinsamen Inhalt zu umschreiben. Jeder Eingeweihte weiß, welche Verfahren zur ‚projektiven‘ Gruppe gehören; aber er unterstellt ihnen kein gemeinsames Konzept (Hörmann, 1982, 178).

Zum andern kann ‚projektiv‘ einen bestimmten **Inhalt** bezeichnen und eine Klasse von Verfahren markieren, denen gemeinsam ist, daß ihr Einsatz den **Vorgang der Projektion** „provoziert“. Dem Probanden wird Reizmaterial vorgelegt, das ihn veranlassen soll, Gedanken und Gefühle zu äußern in Zeichen, in Handlungen, in Worten - generell: in ‚Gestalten‘. Diese ‚Gestalten‘, so lautet die Deutehypothese, ermöglichen es dem Untersucher, die Gedanken und Gefühle des Probanden zu erkennen und aus ihnen auf die ‚Persönlichkeit‘ zu schließen.

Es wird also angenommen, daß der Proband ‚Eigenarten‘ seiner Persönlichkeit in die ‚Gestalten‘ hineinprojiziert und der Untersucher sie darin ‚entdecken‘ kann (Leichner, 1983, 83).

Was besagt nach dieser Konzeption der Ausdruck ‚projektiv‘?

- Im Sinne der **Psychoanalyse** ist Projektion ein unbewußter Vorgang. „Die Projektion erscheint immer als eine Abwehr, in der das Subjekt dem anderen - Person oder Sache - Qualitäten, Gefühle, Wünsche, die es ablehnt oder in sich selbst verleugnet, unterstellt“ (Laplanche & Pontalis, 1977, 403).

Beispiel: *Person A behauptet: „Person B haßt mich.“ Wie jedoch eine Anamnese erkennen läßt, ist es in Wirklichkeit die Person A, welche Person B haßt. Aber das Erlebnis „Ich hasse B.“ widerspricht dem Selbst-Ideal von A, das Ich wehrt das Erlebnis ab und wandelt es in eine Projektion, die besagt: „Person B haßt mich.“*

Nach Freud tritt Projektion nicht nur bei Neurotikern auf, sondern auch bei Normalen, immer aber als Abwehrvorgang.

- Eine **allgemeinere Bedeutung** von Projektion besagt: „Das Subjekt nimmt das umgebende Milieu wahr und antwortet darauf je nach seinen eigenen Interessen, Fähigkeiten, Gewohnheiten, beständigen oder momentanen affektiven Zuständen, Erwartungen, Wünschen etc.“ (Laplanche & Pontalis, 1972, 401).

Beispiel: *Probanden, die mehrere Stunden gehungert hatten, wurden gebeten, Geschichten zu TAT-Tafeln zu erzählen. Es ergab sich: Sie erwähnten häufiger Nahrungsmittel als Personen, die keinen Hunger hatten (Atkinson & McClelland, 1948).*

Dieses Konzept bezeichnet den Vorgang, daß ein Subjekt ‚individuelle Innenwelt in Außenwelt abbildet‘. Es kommt dem Begriff nahe, mit dem Frank (1939) die ‚projektiven‘ Verfahren charakterisiert hat - jener Mann, dem das Urheberrecht zugesprochen wird (nicht für die Prägung, wohl aber) für die weite Verbreitung des Begriffs ‚**projektive** Methoden‘.

Frage einer Einteilung

‚Projektion‘ wird höchst unterschiedlich eingeteilt, zum Beispiel nach Murstein und Pryer (1959, 56, 353-374):

- **Klassische Projektion:** Unakzeptierte Impulse werden abgewehrt, es handelt sich um einen unbewußten Abwehrvorgang (Projektion im Sinne Freuds). Ein **Beispiel** sei wiederholt: A behauptet „B haßt mich.“ In Wirklichkeit haßt A die Person B. Diese ‚Selbsterkenntnis‘ wandelt A jedoch in die Projektion um: „Person B haßt mich.“
- **Autistische Projektion:** Eigene Bedürfnisse färben ‚äußere Wahrnehmungen‘ mit. Die Konturen der äußeren ‚Reize‘ werden auf eigene Wünsche abgestimmt. **Beispiel:** Hungernde Probanden erkennen in ‚projektiven Reizmustern‘ Nahrungsmittel.
- **Rationalisierende Projektion:** Probanden ‚verlegen‘ eigene Wünsche nach außen, bemerken aber ihre eigene Projektion und rechtfertigen sie. **Beispiel:** Person A kritisiert Person B, bemerkt jedoch, daß sie nur ihre Abneigung gegen B abreagiert, und rechtfertigt sich, indem sie bei Person B Gründe für ihre Kritik sucht - ein Exempel für ‚Rationalisierung‘.
- **Attributive Projektion:** Eigene Motive, Gefühle, Verhaltensweisen werden anderen Personen zugeschrieben. Entscheidend ist: Der Aspekt des Abwehrmechanismus bleibt außer Betracht. **Beispiel:** Person A neigt dazu, bei ihrem Partner B ‚Eigenschaften zu erkennen‘, die sie sich auch selber zuspricht (A ist „weich“, B erscheint der Person A ebenfalls als ‚weich‘).

Zur Kennzeichnung ‚projektiver‘ Verfahren sei das Adjektiv in dem allgemeinen Sinne der ‚attributiven Projektion‘ verstanden: Der Proband verlegt eigene Vorstellungen, Wünsche, Bedürfnisse in jene ‚Gestalten‘, die er entwirft, wenn er ‚Kleckse deutet‘ oder ‚Geschichten erzählt‘, wenn er ‚Bilder malt‘ oder aus vorgelegten Spielmaterialeien ‚Szenen formt‘. Die weitere Interpretation des Adjektivs ‚projektiv‘ bleibt offen.

11.2 Klassifikation projektiver Verfahren

Wie bei Tests und Fragebögen gibt es auch bei den projektiven Verfahren keine unumstrittene Einteilung. Wir folgen der Dreiteilung, die Brickenkamp (1975, 13) oder Groffmann und Michel (1982 b) vorgeben. Unterschieden werden

- Formdeutungsverfahren,
- verbal-thematische Verfahren sowie
- zeichnerische und gestalterische Verfahren.

Formdeutungsverfahren fordern vom Probanden, daß er unstrukturiertes Reizmaterial - zufällige, aber symmetrische Klecksgebilde - betrachtet und dann berichtet, was er ‚wahrnimmt‘. Ein typisches Beispiel ist das „Wahrnehmungsexperiment“ von Rorschach (1972).

Verbal-thematische Verfahren verlangen vom Probanden, zu mehrdeutigen Bildern Geschichten zu erzählen. In den Geschichten versucht der Anwender ‚Themata‘ zu erkennen, welche die Persönlichkeit des Probanden charakterisieren. Ein Prototyp ist der „Thematische Apperzeptions-Test“ (TAT) von Murray (1943).

Zeichnerische und gestalterische Verfahren stellen dem Probanden die Aufgabe, vorgegebene Themen zeichnerisch darzustellen (z. B. einen Menschen, einen Baum, ein Haus) oder aus vorgegebenen Materialien (z.B. aus Puppen, Tieren, Bäumen, Farbplättchen) etwas zu gestalten. Vertreter dieser Verfahrensklasse sind der „Baum-Test“ von Koch (1972) oder der „Scenotest“ von Staabs (1964).

Aus dieser Kennzeichnung dürfte hervorgehen, daß in den drei Verfahrensklassen der Vorgang der ‚Projektion‘ unterschiedlich verläuft. ‚Mit den Händen‘ Gestalten zu bilden (in Zeichnungen oder mit Puppen) ist etwas anderes als ‚Geschichten zu erzählen‘ (im TAT) oder ‚Wahrnehmungen wiederzugeben‘ (im ‚Rorschach‘).

11.3 Probleme projektiver Verfahren

Wer projektive Verfahren zur Erfassung vielfältiger Verhaltensprozesse verwendet, benutzt als Deutehilfe in der Regel den Hypothesenvorrat psychodynamischer Persönlichkeitstheorien.

Es gibt auch andere theoretische Grundannahmen: die Forschung zur sozialen Wahrnehmung (social perception) oder die Theorie des Adaptationsniveaus von Helson (Kornadt & Zumkley, 1982, 265-271). Diese Ansätze seien nur erwähnt, nicht skizziert.

*Alle Klassen projektiver Verfahren werfen **Probleme** auf:*

- Der Anreiz (Klecks, Bild, Spielmaterial) bleibt mehrdeutig, ist nicht ‚wohldefiniert‘.
- Die **Reaktionen** („Wahrnehmungen“, „Geschichten“, „Zeichnungen“ oder „Szenarios“) lassen sich oft nicht eindeutig den Auswertungskategorien zuordnen; sie liefern somit auch keine eindeutigen Indikatoren für Persönlichkeitsmerkmale.

Damit sind **Unklarheiten** markiert, die sich beziehen auf

- den *Aufforderungscharakter* des Reizmaterials (Durchführungsobjektivität);
- den *Auswertungsprozeß*, der nur schwer zu einem intersubjektiven Konsens zu führen ist (Auswerter-Objektivität);
- das *Abgrenzungsproblem*, das darin besteht, „fehlerhafte“ und „wahre“ Anteile in den Äußerungen des Probanden zu bestimmen (Reliabilität);
- den *Interpretationsprozeß*, der den Schluß vom Indikator auf das indizierte Merkmal betrifft (Validität).

Ein Teil dieser Probleme stellt sich nur, wenn man die projektiven Verfahren von der klassischen Testtheorie her bewertet. Indessen eignet sich diese zu einer Bewertung nur in begrenztem Maße, weil sie eher zur Messung (relativ) stabiler Eigenschaften (traits) als zur Erfassung von Prozessen entworfen ist (Kornadt & Zumkley, 1982, 332).

Ein Teil der Probleme stellt sich aber auch unabhängig von der klassischen Testtheorie, vor allem das der Relation von Index und indiziertem Merkmal. In dieser Frage muß sich der Untersucher auch dann Gewißheit verschaffen, wenn er einen anderen meßtheoretischen Rahmen als die klassische Testtheorie wählt (vgl. Wittkowski, 1996).

11.4 Beitrag projektiver Verfahren zur Diagnostik

Sicher ist es möglich, projektive Verfahren zu nutzen als Instrumente einer Intervention. (So schlagen Revers & Allesch (1985) vor, ihre Variante des TAT als Mittel einer Therapie zu verwenden: als Instrument, den Probanden mit seiner Biographie zu konfrontieren.) Zentrieren dürfte sich ihre Verwendung jedoch in der Diagnostik.

Das Problem der Zuordnung von Index und Indiziertem sollte jedem Anwender projektiver Verfahren Vorsicht auferlegen - Zurückhaltung gegen jede ihrer Interpretationen. Kein Entscheidungsvorschlag (etwa in Beratung oder Begutachtung) sollte allein auf projektiven Verfahren gründen.

Aber die **Vorteile**, die sie bieten, kann er nutzen:

- ‚Projektive Aussagen‘ (Deutungen, Geschichten, Zeichnungen) können ihn auf Probleme aufmerksam machen, die den Probanden belasten, die er jedoch nicht nennen kann oder nicht nennen will. Sie können ihn ‚auf die Suche‘ schicken und ihn ‚anleiten‘, das signalisierte Problem mithilfe anderer Verfahren zu erkunden. Solche **‚Heurismen‘** bereitzustellen, darin dürfte ein charakteristischer Beitrag projektiver Verfahren liegen.
- Diesem ‚heuristischen Beitrag‘ ist es dienlich, daß ein Proband kaum einschätzen kann, was projektive Verfahren diagnostisch erschließen sollen. Sie sind für ihn **weniger durchschaubar** als etwa Persönlichkeitsinventare, darum auch weniger Verfälschbar.

- Deutlicher als in Verhaltensbeobachtung, als in Tests oder Fragebögen können sich in den ‚projektiven Gestalten‘ (Wahrnehmungen, Geschichten, Szenen) auch die ***Genese oder die situationalen Momente eines Problems*** abzeichnen, z. B. die Entstehung und der Verlauf eines Partnerschaftskonfliktes.

Ermöglicht wird diese Art der Schilderung durch eine besondere Eigenart: Projektive Verfahren nötigen den Probanden zu ***kreativem Tun***, etwa Geschichten zu erzählen, Kleckse zu deuten, Figuren zu Szenen zu gruppieren.

„Die psychometrischen Tests begreifen das Individuum eher als Träger . . . von Fähigkeiten und Fertigkeiten, während die Diagnostik durch projektive Verfahren mehr das potentiell Kreative von Individuen, . . . aber auch das Zusammenwirken von inneren und äußeren gleichgewichtserhaltenden und destabilisierenden Kräften betont“ (Spitznagel, 1990, 409-410).

Zur sprachlichen, motorischen, kreativen Kompetenz

Formdeutungsverfahren und verbal-thematische Verfahren setzen ein gewisses (kaum definierbares Maß) an sprachlicher Kompetenz voraus, nämlich die Fähigkeit, das ‚Gesehene‘ sprachlich zu formulieren (beim Rorschach) oder die dargestellte Szene ‚phantasievoll‘ auszumalen (beim TAT). - Zeichnerische und gestalterische Verfahren setzen ein gewisses Maß an motorischer ‚Geschicklichkeit‘ voraus, beim Zeichnen ebenso wie beim Gruppieren von Puppen. - In der sprachlichen und motorischen Kompetenz unterscheiden sich die Probanden.

Was projektive Verfahren jedoch erfassen sollen, ist nicht der unterschiedliche Grad an sprachlicher, motorischer, kreativer Kompetenz. Erschließen sollen sie Verhaltensanteile, denen die ‚projektiven Antworten‘ entspringen: die Impulse, Vorstellungen, Wünsche, Ängste, Motivationsstrukturen des Probanden.

11.5 Darstellung von drei Klassen projektiver Verfahren

Das weitere Kapitel skizziert drei Klassen projektiver Verfahren einzeln:

- Formdeutungsverfahren (11.5.1),
- verbal-thematische Verfahren (11.5.2),
- zeichnerische und gestalterische Verfahren (11.5.3).

11.5.1 Formdeutungsverfahren

Formdeutungsverfahren legen dem Probanden bedeutungslose oder bedeutungsarme Gebilde vor (etwa „Kleckse“) und fordern ihn auf, in Worte zu kleiden, was er in den Gebilden „sieht“. Dieses „Sehen“ als kognitiver Prozeß ist ein

Ergebnis der aktiven Gestaltung des Probanden; insofern enthalten die Deutungen ‚Eigenarten‘ seiner Person. Diese sollen aus den Antworten erschlossen werden, in erster Linie aus ihren formalen Elementen, erst in zweiter Linie aus ihren Inhalten (Spitznagel, 1982 b).

Das bekannteste Beispiel ist das Verfahren, das Rorschach im Jahre 1921 als „wahrnehmungsdiagnostisches Experiment (Deutemassen von Zufallsformen)“ eingeführt hat. - Kasten 11-1 nennt Varianten.

Kasten 11-1: Varianten von Formdeutungsverfahren

- Schüler, Nachahmer, Kritiker haben Rorschachs Konzept ergänzt, erweitert, verändert, zum Beispiel:
- Von **Behn-Eschenburg** (1952) liegen zehn Parallel-Tafeln zu Rorschachs Originalserie vor (BERO-Test: siehe Zulliger, 1952).
 - Fuchs, Ch. (1958) hat desgleichen zehn Paralleltafeln zur Originalserie veröffentlicht (Fuchs-Rorschach-Test: FURO-Test).
 - **Holtzman** (1961, 1972) hat es unternommen, ein Formdeutverfahren nach psychometrischem Vorbild zu konzipieren: zwei Parallelformen zu je 45 Tafeln, die in Analogie zu einem klassischen Test ausgewertet werden sollen (Holtzman Inkblot Technique: HIT).
 - **Klopfer und Davidson** (1942, 1974) haben genauere Regeln zur Vorgabe, Auswertung und Interpretation der zehn Rorschach-Originaltafeln vorgelegt; „alte“ Rorschach-Kürzel wurden ins Englische übertragen
 - **Zulliger** (1955) hat drei Bilder in Entsprechung zu den Rorschach-Tafeln entworfen, die sich von Diapositiven auf Leinwand projizieren lassen, so daß sie als Gruppenverfahren anwendbar sind (Diapositiv-Z-Test: Dia-Z-Test).

Zum Formdeutungsverfahren nach Rorschach

Die Auswertung von Formdeute-Protokollen erfordert eine Vielzahl von Schritten. Hier seien nur genannt:

1. die Signierung der einzelnen Antworten,
2. ihre Zusammenfassung (Verrechnung) in einem *Psychogramm*,
3. ihre *Interpretation* in einem fortlaufenden Text.

Zu 1.: Die **Signierung** besteht darin, jede einzelne Antwort in ein Kürzelsystem zu transskribieren. Verwiesen sei auf drei Aspekte, die nach Rorschach Einteilungskriterien liefern:

- **Lokalisation:** Welchen Klecksteil hat der Proband einbezogen, den Gesamtklecks oder nur einen Teil?
- **Determinantenbestimmung:** Was an dem Klecks hat den Probanden zur Antwort gebracht: die „Farbe“ oder die „Form“ oder auch eine „wahrgenommene Bewegung“?
- **Inhalt:** Welchen Inhalt stellt die Wahrnehmung dar, etwa ein Tier oder einen Mensch oder eine Pflanze?

Die Antwortvielfalt wird also auf wenige Klassen reduziert, nur drei wurden hier genannt.

Zu 2.: Die Signierungen der Einzelantworten werden gezählt und in einer Übersicht zusammengefaßt, sie ergeben ein **Psychogramm**. Meist werden dabei mehrere Kürzel zu einzelnen Gruppen geordnet (bei Klopfer & Davidson [1974] zu sogenannten ‚Proportionen‘).

Zu 3.: Die **Interpretation** fußt auf dem Psychogramm und den Deutehypothesen, die den Kürzeln oder Kürzelgruppen zugeordnet sind. Sie „stutzen“ Aussagen, die ein breites Band von Merkmalen betreffen, beispielsweise

- Intelligenz,
- kognitiven Zugangsstil zu Sachverhalten (Erfassungstyp),
- Reichtum, Originalität, Konventionalität der Antworten,
- Leistungsverhalten (Bereitschaft, Anspruch, Kapazität),
- Erlebnisrichtung (Erlebnistyp: introversiv, ambivertiert, extrativ),
- Kontrolle der Emotionen,
- usw.

Schon diese - simplifizierende - Skizze der Anwendung und Auswertung rechtfertigt den Ratschlag, hohe Vorsicht im Gebrauch der Formdeuteverfahren walten zu lassen.

Für keinen der drei Schritte (Signierung, Verrechnung, Interpretation) liegen so eindeutige Zuordnungsregeln vor,

- daß zwischen Auswertem Konvergenz selbstverständlich ist (Objektivität),
- daß Replikationen identisch ausfallen (Retest-Reliabilität) und
- daß Antworten immer als Indikatoren für ‚dieselben‘ Merkmale stehen (Validität).

Wegen der Vielzahl ungelöster Probleme sollten Formdeuteverfahren nur Hilfsdienste übernehmen, zum Beispiel zur Generierung von Heurismen.

Beispiele für Heurismen aus Formdeuteverfahren

1. Ein Proband widerruft immer wieder seine Antworten im Rorschach, ersetzt gegebene Deutungen durch neue Deutungen. Ein solches Verhalten könnte den Untersucher zu dem Schluß verleiten, der Proband sei ein „unentschlüsselter Mensch“ - der Schluß dürfte voreilig sein.
Sehr wohl kann ein solches Verhalten den Untersucher veranlassen, zu erkunden, ob der Proband Schwierigkeiten habe, Entscheidungen zu treffen: in der Berufswahl, in der Wahl von Freunden, in der zeitlichen Planung eigener Arbeiten usw.
2. Ein Proband bringt keine Sexualdeutung, auch wenn eine Klecksform eine „sexuelle Wahrnehmung“ nahelegt. Wieder läßt sich nicht einfach folgern, der Proband habe sexuelle Probleme.

Sehr wohl kann eine solche ‚Zurückhaltung‘ den Untersucher veranlassen, die Thematik Sexualität im Gespräch zu behandeln - sofern die Fragestellung das Thema einschließt.

HINWEIS: *Einem Anwender ist dringend zu raten, Formdeutungsverfahren nur dann einzusetzen, wenn er eine gründliche und kritische Schulung durchlaufen hat.*

11.5.2 Verbal-thematische Verfahren

Verbal-thematische Verfahren bezeichnen eine Gruppe diagnostischer Instrumente, die darauf angelegt sind, „Persönlichkeitseigenarten aus Geschichten zu erschließen, die der Proband zu bestimmten Bildern erzählen soll... Im allgemeinen ist das Ziel dieser Verfahren, aus der Auffassung und Bearbeitung des Bildthemas einen Aufschluß über die inhaltliche, gegenstandsbezogene Seite der Persönlichkeitsdynamik zu erhalten“ (Kornadt & Zumkley, 1982, 259).

Verbal-thematische Verfahren verlangen, daß der Proband zu Bildern, die ihm vorgelegt werden (und die meist mehrdeutig gestaltet sind), Geschichten erfindet. In ihnen will der Untersucher ‚Themen‘ ausmachen, die für den Erzähler charakteristisch sind. Sie resultieren aus den ‚Bedürfnissen‘, die im Erzähler entstehen (needs), und aus ‚Eindrücken‘, die er von der Umwelt empfängt (press).

Die letzten Sätze passen am besten auf den Hauptrepräsentanten dieser Verfahrensklasse, auf den ‚Thematischen Apperzeptions-Test‘ von Murray (TAT: 1943). Bevor er vorgestellt wird, seien in Kasten 11-2 einige Abwandlungen genannt.

Kasten 11-2: Varianten verbal-thematischer Verfahren

Es seien einige Beispiele genannt, die sich als Abwandlungen des Thematischen Apperzeptions-Tests verstehen lassen:

- **Lennepe (1948), Lennepe und Houwink (1958)** veröffentlichten den ‚Vier-Bilder-Test‘ (Four Picture Test): Zu vier Bildern, die keinen Zusammenhang bilden, soll der Proband eine fortlaufende Geschichte erzählen.
- **Bellak und Bellak (1949)** publizierten zehn Tafeln, auf denen Tiere als Figuren auftreten (Childrens's Apperception Test: CAT). Die Tiere sollen Kindern das Erzählen erleichtern.
- **Phillipson (1955)** entwarf aus psychoanalytischer Konzeption drei Serien zu je vier Bildern (Object Relations Technique: ORT); die zwölf Tafeln stellen Einer-, Zweier- und Gruppensituationen vor, die einen ‚Erzähler‘ anregen sollen, seine psychosexuelle Entwicklung darzustellen.
- **Soloman und Starr (1968)** gaben auf zwölf Tafeln Schulszenen vor, zu denen ‚Schüler‘ als Probanden Geschichten entwerfen sollen (School Apperception Method: SAM).
- **Bellak und Bellak (1973)** veranschaulichte auf sechzehn Tafeln Szenen aus dem Leben älterer Menschen (Senior Apperception Technique: SAT).

- **Revers und Allesch** (1985) verfaßten zum Thematischen Gestaltungstest TGT-S das Handbuch; sie versuchten, Auswertung und Interpretation stärker in die Biographie des Probanden einzuordnen. Das Diagnostikum TGT-S soll dazu dienen, einen Probanden mit seiner Biographie zu konfrontieren - ein Übergang von der Diagnostik zur Intervention. (Vgl. die Kontroverse zum TGT: Steck, 1989, 1991, und Tent, 1991).
- **Rosenzweig** (1945, 1948) konzipierte mit dem Picture Frustration Test (PFT) ein *semi-projektives* Verfahren: 24 Zeichnungen geben ‚Begegnungen‘ wieder, in denen eine Person Frustrationen erleidet. Aus der Situation der frustrierten Person heraus soll der Proband eine Antwort formulieren. Die Gesamtheit der Antworten soll einen Schluß ermöglichen auf die Art und die Richtung der Aggression des Antwortgebers. - Der PFT liegt als Kinder- und als Erwachsenenserie vor. Normen wurden erstellt in Formen von Quartilwerten.
- **Rauchfleisch** (1979) faßte den Rosenzweig-Picture-Frustration-Test (PFT) neu und adaptierte die Auswertungsregeln an psychometrische Verfahren. Er eichte die Kinder- und die Erwachsenenserie neu.

Zum Thematischen Apperzeptions-Test

Der Prototyp der verbal-thematischen Verfahren, der Thematische Apperzeptions-Test (TAT), sei exemplarisch skizziert. Der Bezeichnung nach ist der TAT ein ‚**Test**‘.

Frage: *Ist der TAT aber wirklich ein Test, also „ein wissenschaftliches Routineverfahren zur Untersuchung eines oder mehrerer empirisch abgrenzbarer Persönlichkeitsmerkmale mit dem Ziel einer möglichst abgestuften Aussage über den relativen Grad der individuellen Merkmalsausprüfung“ - wie die Definition eines Tests nach Lienert-Raatz lautet (1994, I)?*

Der Sache nach läßt sich der TAT nicht in diesem Sinne verstehen, er ist kein psychometrisches Verfahren. Das Wort ‚Test‘ trifft nur in dem allgemeinen Sinne eines Prüfverfahrens zu. *Hier bezeichnet - nach der Sprachregelung dieses Buches - der Titel ‚Test‘ einen Namen* (vgl. S. 27).

Der TAT hat zu tun mit **Apperzeptionen**, also mit Wahrnehmungsprozessen: Was der Proband auf den ihm vorgelegten Bildtafeln „wahrnimmt“, soll ihn zu einer Gestaltung geschlossener Geschichten anregen.

Aus den Geschichten sollen **Themata** erschlossen werden: Motivstrukturen, die zwei Quellen entspringen, den Bedürfnissen, die im Probanden selber vorherrschen, und den Einflüssen, welche die Umwelt auf ihn ausübt (needs and press: Murray, 1938, 1943). Diese Themata sollen nach Murray die Individualität des Probanden erfaßbar und kenntlich machen (unity themes).

Das **Material** besteht aus einunddreißig Tafeln, *dreißig* bieten Schwarz-Weiß-Bilder, *eine* Tafel ist ein leeres weißes Blatt. Die Tafeln sind für unterschiedliche Gruppen vorgesehen: Elf Bilder sind für alle Altersstufen und für beide Geschlechter bestimmt, von den anderen Bildern Teilmenge jeweils nur für

Erwachsene (getrennt nach Männern und Frauen) oder nur für Kinder (getrennt nach Jungen und Mädchen).

Dem Probanden sollen zwei Serien zu verschiedenen Zeitpunkten vorgegeben werden, jede Serie zu je zehn Tafeln. Die erste Serie, Nr. 1-10, bietet realistischere Bilder, die zweite Serie, Nr. 11-20 bizarrere, phantastischere Bilder. (Auf der Rückseite tragen die Bilder Zahlen, welche die Zuordnung zu den beiden Serien ermöglichen.)

Zu den Bildern läßt Murray Geschichten erzählen mit der Instruktion: „Erzählen Sie eine Geschichte zu diesem Bild, die möglichst dramatisch ist. Berichten Sie, wie es zu dieser Szene kam, was jetzt vor sich geht und wie die Geschichte ausgeht.“

Auf zwei Punkte der Instruktion sei verwiesen: auf die Aufforderung, die Geschichten dramatisch zu erzählen, und auf den Hinweis, die Geschichten in drei Zeitdimensionen zu gestalten (Was war? Was ist? Was wird sein?).

- Die *Forderung nach dramatischer Gestaltung* soll den Probanden davon ablenken, beim Erzählen besondere ‚Leistungen‘ erbringen zu wollen.
- Der Hinweis auf die *drei Zeitdimensionen* (Vergangenheit, Gegenwart, Zukunft) soll Anhaltspunkte geben für den Verlauf der Dramatik, zugleich soll sie den Erzählenden in einen weitgespannten Zeitrahmen setzen.

Zur **Auswertung** des TAT sei nur soviel gesagt, daß sie bestimmte theoretische Rahmenkonzepte voraussetzt - bei den meisten Anwendern die sogenannte *Projektions-* und die *Identifikationshypothese*. Über beide ist eine lebhaft Diskussion geführt worden (Kornadt & Zumkley, 1982, 281; Lindzey, 1967).

- Die **Projektionshypothese** wurde schon besprochen: Es wird angenommen, daß der Proband in den Geschichten eigene Vorstellungen, Wünsche, Impulse darstellt. (Es wird nicht angenommen, daß die Geschichte reales Verhalten des Erzählers wiedergibt.)
- Die **Identifikationshypothese** besagt, daß die Geschichten eine oder mehrere Figuren enthalten, mit denen sich der Erzähler in besonderer Weise identifiziert. Diese Annahme sei kurz erläutert:
 - ⇒ Identifikation leitet sich als Wort ab vom lateinischen Pronomen ‚idem‘, welches ‚dasselbe‘ bedeutet. Identifikation besagt demnach, daß zwei-erlei Dinge ‚als dasselbe‘ betrachtet, daß sie gleichgesetzt werden, Identifikation bedeutet also ‚Gleichsetzung‘.
 - ⇒ Transitiv bedeutet Identifikation, zwei Dinge als identisch zu betrachten. Wenn jemand eine Person nach einem Bild ‚identifiziert‘, dann werden Mensch und Bild gleichgesetzt. (In diesem Sinne heißt der Personalausweis auch Identifikationskarte.)
 - ⇒ Intransitiv besagt Identifikation, sich selber gleichsetzen mit jemandem oder mit einer Sache. Hierher paßt die Redewendung: „Ich identifiziere mich mit meinem Partner.“
 - ⇒ Wie der Begriff der Projektion, so kommt auch das Konzept der Identifikation aus der Psychoanalyse. Dort bezeichnet Identifikation den

Vorgang, durch den ein menschliches Objekt sich selbst konstituiert: Ein Kind will werden wie sein erstes Liebesobjekt, wie die Mutter, und wird dadurch zu einem eigenen Selbst. In der ödipalen Phase will der Junge werden wie der Vater, er setzt sich von ihm ab, setzt sich ihm entgegen und setzt sich mit ihm gleich, auch hier bezeichnet Identifikation den Weg zur Selbstwerdung.

⇒ In einem abgeleiteten Sinne bezeichnet Identifikation auch den Vorgang, in dem ein Subjekt sich mit ‚Teilen‘ einer anderen Person gleichsetzt. Demgemäß sind sehr viele partielle Identifikationen möglich, dabei brauchen die Identifikationen kein kohärentes System zu bilden.

Resümee: *In einem allgemeinen Sinne bezeichnet Identifikation den Vorgang, bei dem ein Subjekt seine Wünsche, Gefühle, Gedanken auf eine andere Person oder Sache überträgt. Damit nähert sich der Begriff der Identifikation dem der Projektion. Er bezeichnet jenen Teil der Projektion, der sich auf das projizierende Subjekt selber bezieht. Mit Identifikation ist jene Teilmenge der Projektion gemeint, in der ein Proband Wünsche, Hoffnungen, Befürchtungen äußert, die seine eigene Person betreffen.*

Wenden wir die Umschreibung auf den TAT an, dann besagt die ‚Identifikationshypothese‘ in ihrer allgemeineren Bedeutung: Gedanken, Gefühle, Wünsche einer bestimmten Figur in den erzählten Geschichten repräsentieren besonders deutlich Gedanken, Gefühle, Wünsche des Erzählers selber. Diese Figur wird ‚Hauptfigur‘ (hero: Held) genannt.

Murray selbst gibt Regeln vor, die angeben, wie sich die ‚Hauptfigur‘ in den Geschichten erkennen läßt. Aber er sieht auch den Fall vor, daß sich keine Hauptfigur ausmachen läßt. - Kornadt und Zumkley bezweifeln die Möglichkeit, die Identifikationshypothese „als *allgemein* anwendbares Auswertungsprinzip“ anzunehmen (1982, 279).

„Da sich heute . . . kaum noch die Intention findet, im TAT ‚die ganze Persönlichkeit‘ erfassen zu wollen, sondern Aussagen mehr für umschriebene Persönlichkeits- und Motiv-Bereiche gemacht werden . . . ist die Bedeutung der Identifikations-Hypothese als generelles Auswertungsprinzip relativiert“ (Kornadt & Zumkley, 1982, 280).

Es gibt **unterschiedliche Auswertungssysteme** - *quantitative* Methoden (beispielsweise von Murray, 1943) ebenso wie *qualitative* Erschließungswege (beispielsweise von Revers, 1979). In jedem Falle sollte der Auswerter einem Schema folgen, das er rational begründen kann.

HINWEIS: *Werden die Geschichten auf Bildebene ausgewertet, in Analogie zu Träumen oder assoziativen Einfällen, so muß auch eine solche Interpretation - sie erst recht - intersubjektiv begründbar sein. Sie setzt ein eingehendes Training und eine ständige Supervision voraus, sonst verfallt sie in kaum nachvollziehbare Spekulationen.*

Wie bei der Exploration raten wir auch hier **dem Anfänger** zu einer *Auswertung in mehreren Stufen*:

- Der Anfänger sollte zunächst jede *Geschichte einzeln* auswerten, indem er ‚Themen‘ identifiziert.
- Sodann sollte er versuchen, *wiederkehrende Themen* stichwortartig *zusammenzufassen*.
- Schließlich, in bestimmten Fällen, z.B. bei Erstellung eines Gutachtens, sollte er als Gesamtauswertung *einen fortlaufenden Text formulieren*.

Drei Bemerkungen zu TAT-Auswertungen

Erstens, der Auswerter sollte nicht über die Annahme hinausgehen, daß der Proband in seinen Geschichten eine Verhaltensstichprobe seiner Vorstellungen bietet, Assoziationen, zu denen die TAT-Tafeln ihn anregen. Diese Assoziationen können über Verhaltensbereitschaften des Probanden Auskunft geben, nicht über sein tatsächliches Verhalten.

Zweitens, es ist kaum möglich, ein TAT-Protokoll vollständig auszuwerten. Der Untersucher sollte sich klar sein, daß er bei der Auswertung mit Blick auf die Fragestellung eine Auswahl trifft.

Drittens, die TAT-Protokolle sollten zur Formulierung von Heurismen anregen. Sie können helfen, Fragen zu formulieren für andere Verfahren, z.B. für Gespräche. Sie können auch Anhaltspunkte geben für die Interpretation anderer Verfahren, z.B. der Ergebnisse von Fragebogen. In diesem Sinne lassen sich Befunde des TAT an anderen Verfahren ‚validieren‘ (oder ‚invalidieren‘).

HINWEIS: *Wer je einmal einer Gruppe ‚uneingeweihter‘ Zuhörer TAT-Geschichten vorlas und dann versuchte, eine Auswertung mit ihnen zu erarbeiten, der weiß, wie rasch sich die ‚Interpreten‘ von ihrer Phantasie zu weitreichenden, ja ausschweifenden Deutungen hinreißen lassen - die allesamt höchst ‚plausibel‘ erscheinen ...*

Beispiele für Heurismen aus dem TAT

1. Eine Frau erzählt Geschichten, in denen zwei Rivalinnen um denselben Mann kämpfen, aber beide von ihm verlassen werden. Aufgrund dieses Themas sollte der Untersucher nicht etwa schließen, daß die Erzählerin in Auseinandersetzung mit einer Freundin/Rivalin um denselben Mann stehe. Er sollte sich für ein Gespräch höchstens die Frage vorgeben lassen, ob die Erzählerin partnerschaftliche Beziehungsprobleme habe, ob in den zwei Rivalinnen der Geschichte nicht etwa zwei Impulsrichtungen der Erzählerin selber repräsentiert sein könnten: widerstreitende Gefühle in partnerschaftlichen Beziehungen.

2. Ein Mann erzählt Geschichten, in denen Figuren Orgien von Aggressionen feiern. Wiederum sollte der Auswerter es vermeiden, unmittelbare Schlüsse zu ziehen, etwa aus den Beschreibungen ‚aggressives‘ Verhalten des Probanden zu erschließen.

Das Thema könnte ihn aber veranlassen, der Rolle von Aggression im Leben des Probanden nachzugehen, sich dabei auch sein soziales Verhalten auf konkreter Ebene beschreiben zu lassen.

Training und Supervision: Für Durchführung, Auswertung, Interpretation verbal-thematischer Verfahren erhält der Ratschlag höchste Dringlichkeit, diese Instrumente – die Wegweiser sein, aber auch zu Verführern entarten können – nur anzuwenden, wenn ein intensives **Training** vorangegangen ist und kritische **Supervisoren** den Anwender ständig begleiten.

11.5.3 Zeichnerische und gestalterische Verfahren

Bei zeichnerischen und gestalterischen Verfahren wird der Proband gebeten, ein vorgegebenes Thema oder einen eigenen Einfall zu gestalten (Sehringer, 1982).

Zeichnerische Verfahren verlangen, daß der Proband eine Zeichnung erstelle, meist ist das Thema vorgegeben. Die Instruktion für ein Kind könnte lauten: ‚Zeichne einen Menschen! Zeichne einen Baum! Stelle die Mitglieder deiner Familie als Tiere dar!‘

Zeichnerische Verfahren, die kein Thema vorgeben, sind die Ausnahme, aber es gibt sie, beispielsweise den Wartegg-Zeichen-Test (WZT: Wartegg, 1939, 1953).

Zwei Beispiele:

1. Beim **Baum-Test** (Koch, 1972) soll der Proband auf weißem Papier einen Baum zeichnen und ihn nach Belieben ausgestalten. Zur Auswertung legt der Autor eigene Analysen von Zeichnungen vor und gibt Hinweise zur Deutung einzelner Merkmale. „Er sieht den Baum als idealen Projektions-träger an, da der Baum zu den ältesten Symbolen der Menschheit überhaupt gehört“ (Brickenkamp, 1975, 539). Das Handbuch bietet Häufigkeitstabellen für Einzelmerkmale an, diese ‚Normen‘ gelten für Altersgruppen von 6 bis 16 Jahren.
2. Bei dem Verfahren **Familie in Tieren** (Brem-Gräser 1975) soll der Proband – meist ein Kind – die Mitglieder seiner Familie in Tieren darstellen, auch sich selber. Für die Deutung der Tiere gibt die Autorin Hilfen, gewonnen aus Befragungen, in denen ‚Bedeutungen‘ erhoben wurden, die mit Tieren assoziiert werden.

Gestalterische Verfahren fordern den Probanden auf, aus vorgegebenen Materialien (etwa Puppen, Tieren, Bäumen, Häusern) eigenen Vorstellungen „Ausdruck“ zu verleihen oder mit ihnen ein bestimmtes Thema zu gestalten.

Für die Auswertung zeichnerischer und gestalterischer Verfahren gilt - es sei wiederholt - der Grundsatz: Nicht die Kunstfertigkeit wird bewertet, nicht zeichnerische oder gestalterische Tüchtigkeit, sondern allein der ‚Ausdruck‘ - das, was der Proband von sich in die Gestalten ‚hineinprojizieren‘ will.

Beispiel: Ein bekanntes gestalterisches Verfahren ist der **Steno-Test** (von Staabs, 1964): Dem Probanden, in der Regel einem Kind wird Spielmaterial angeboten, etwa Puppen, Tiere, Bäume, Blumen, Fahrzeuge, Bausteine usw. In einer Miniaturwelt soll er Vorstellungen, Affekte, Konflikte abbilden, die er in der ‚großen‘ Welt erlebt. Zum Abschluß soll er selber schildern, was er dargestellt hat. Der Untersucher soll die dargestellte Szene fotografieren oder auf einem von der Autorin entwickelten Protokollbogen beschreiben. Die Auswertung setzt die Kenntnis tiefenpsychologischer Symbolik voraus.

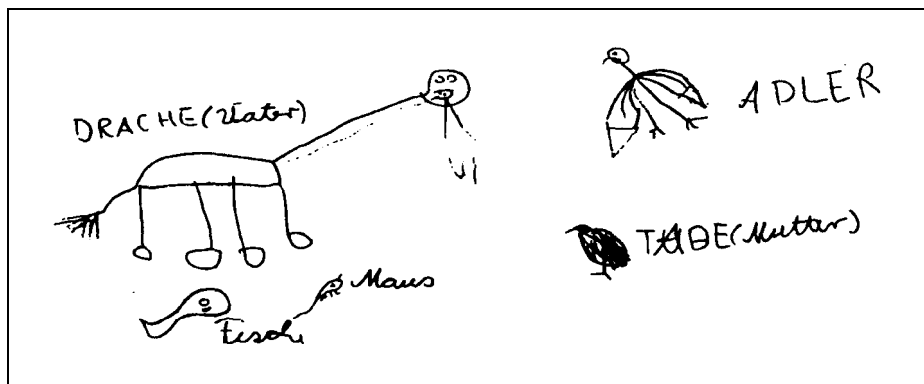
HINWEIS: Wenn die zeichnerischen und gestalterischen Verfahren zum Verständnis des Probanden beitragen sollen, setzen sie ein umfangreiches **Training** und eine immerwährende ‚kollegiale Konsultation‘ voraus, genau wie ‚Rorschach‘ oder TAT.

Ein Beispiel

Kasten 11-3 bringt ein Bild, das - *Anamnese oder Exploration zu einem Fall vorausgesetzt* - einen innerfamiliären Konflikt **veranschaulichen** kann. Es geht um das Verfahren „Familie in Tieren“.

Kasten 11-3:
Beispiel für ein zeichnerisches Verfahren/ „Familie in Tieren“

Siehe Kommentar unter Kasten 11-3!



Kommentar zu dem Bild in Kasten 11-3

Hinweise zur familiären Situation:

- Proband ist Günter, zehn Jahre alt.

Seine *Familie* besteht aus Vater (Diplom-Ingenieur), Mutter (Medizinisch Technische Assistentin) und drei Kindern: Günter (zehn Jahre) Jennifer (vier Jahre) und Dorothee (drei Jahre).

- Die *Eltern* leben und wohnen getrennt, sind aber nicht geschieden. Die Mutter hat die Kinder in ihrem Haushalt, ohne allerdings allein das Sorgerecht zu haben. Der Vater ‚mische‘ sich in die Erziehung vor allem dann ein, wenn es um Fragen der Schullaufbahn gehe: so die Auskunft der Mutter.
- *Konflikt*: Günter möchte von der Grundschule zur Realschule wechseln, die Mutter stimmt zu. Der Vater wünscht, daß er „ein Gymnasium besucht und ordentlich Latein lernt“.

Angaben von Günter zu seinem Bild:

- Adler? „Bin ich selber. Sonst kann ich nichts dazu sagen.“
- Taube? „Ist meine Mutter, sie ist immer gut zu mir, immer freundlich, nein, nicht immer.“
- Drache? „Ist mein Vater. Der ist viel weg, ist nie zuhause, nachts auch, und er schimpft viel.“
- Fisch? „Die Jennifer, die ist lebendig wie ein Fisch im Wasser, ist auch ganz schön stark für ihr Alter.“
- Maus? „Die Dorothee ist noch ziemlich klein. Ziemlich lieb ist sie gar nicht, doch, oft ist sie lieb.“

Hinweise zur Interpretation:

- Es werden Tiere eingeführt, die keinen gemeinsamen ‚Wohnraum‘ einnehmen
- Der Adler ist am höchsten plaziert, er blickt mit der Taube in die gleiche Richtung: gegen den Drachen.
- Nur Adler und Taube sind von derselben Tierart: Vogel.
- Der Drache, die umfangreichste Figur, speit Feuer in die Mitte der anderen Tiergruppe.
- Beherrschendes Gegenüber: Drache und Adler, beide Flugtiere, aber gegensätzlicher Art (‚edel‘, ‚giftig‘).
- Klare Über- und Unterordnung: Drache und Adler ‚oben‘, Taube, Maus und Fisch ‚unten‘.

11.6 Zusammenfassung zu Kapitel 11

Es ist schwierig, die Rolle projektiver Verfahren im diagnostischen Prozeß angemessen einzuschätzen. Schulmeinungen beeinflussen das Urteil. Dies gilt für alle drei Klassen, die skizziert wurden: für Formdeutungsverfahren ebenso wie für verbal-thematische und für zeichnerische oder gestalterische Verfahren.

Die meiste Akzeptanz dürfte heute eine Wertung finden, die besagt: Angaben aus projektiven Verfahren können als Heuristiken für den Einsatz anderer Verfahren dienen, etwa Gespräch oder Verhaltensbeobachtung.

Die Chance projektiver Verfahren liegt darin, Erlebens- und Verhaltensanteile manifest zu machen, die der Proband bewußt nicht benennen kann (oder will), die ihn aber zu seiner diagnostischen Anfrage mitveranlaßt haben.

Die Grenzen projektiver Verfahren markiert die Forderung, zwischen ‚projektivem‘ Index und indiziertem Merkmal eine valide Beziehung anzugeben. Die Auswertung bleibt bislang weitgehend auf Deutehypothesen angewiesen.

11.7 Kontrollfragen zu Kapitel 11

- Definition von „projektiv“, „Projektion“, „Identifikation“.
- Klassifikation projektiver Verfahren.
- Spezieller Beitrag projektiver Verfahren zur Diagnostik.
- Spezielle Gefahrenquellen projektiver Verfahren.

Teil IV

Einzelaspekte integrativer Diagnostik

Dieser Grundriß der Diagnostik orientiert sich an der diagnostisch-interventiven Situation. Viele Aspekte, die in dieser Situation erscheinen, werden in der Behandlung vieler Fragestellungen miterwähnt, aber nicht weiter skizziert.

Teil IV soll solche Einzelaspekte hervorheben. Wir besprechen folgende Themen:

- Ethisch-juristische Determinanten (Kap. 12),
- Klassifikation und Selektion (Kap. 13),
- Statistische und Klinische Urteilsbildung (Kap. 14),
- Drei Ansätze für Diagnostik und Intervention (Kap. 15),
- Erfolgskontrolle (Kap. 16),
- Nutzenschätzung/Entscheidungstheorie (Kap. 17),
- Computergestützte Diagnostik (Kap. 18).

Ethisch-juristische Determinanten von Diagnostik und Intervention

Immer ist der Diagnostiker in einem finalen, sozialen, in einem ethisch-juristischen Kontext tätig. Kapitel 1 hat diese Einbindung erwähnt (S. 8), Kapitel 19 und 21 werden wieder darauf eingehen (S. 415 und 441).

Ethische und juristische Ansprüche sind der diagnostischen Arbeit immanent, sofern sie zu tun hat mit der Eigenständigkeit des Probanden und seiner Beziehung zu anderen Personen, also mit Selbst- und Fremdbestimmung, mit Freiheit und Personenwürde. Damit begegnet diagnostische Arbeit einer ‚Solensordnung‘ des Handelns, die zwar höchst unterschiedlich begründet, von der ‚Gesellschaft‘ aber als Ausdruck allgemein akzeptierter Grundrechte verstanden wird.

Nur zwei ‚Stellen‘ seien genannt, an denen der diagnostische Prozeß solche ethisch-juristischen Imperative gleichsam ‚faßbar‘ macht:

- Diagnostik schließt in vielen Fällen *Kenntnis sehrpersönlicher; ja intimer Suchverhalte* ein. Solche persönlichen Inhalte gehören der Sphäre von Selbstbestimmung an. Über sie darf der Untersucher darum auch nur in der Absicht verfügen, in der sie ihm anvertraut wurden. Die Selbstenthüllung des Probanden und die Kenntnisnahme des Untersuchers vollziehen sich in einem vorgegebenen ethisch-juristischen Kontext.
- Intervention, etwa eine Therapie, kann den Lebensrahmen eines Probanden erheblich beeinflussen (seine Berufswahl, seine Partnerwahl, seine Einstellung zur Gesellschaft). Damit ist wieder eine Sphäre berührt, in der sich Proband und Psychologe gemeinsam bewegen, der Bereich der *Selbst- und Fremdbestimmung*.

Zu Einzelpflichten seien nur Stichworte genannt:

- Respekt vor der *Personwürde* des Probanden während der Untersuchung (keine Demütigung, keine Nötigung);
- Unterrichtung des Probanden über *Untersuchungsziel* und Weitergabe von Ergebnissen;
- Schutz der *Privat- und Intimsphäre* des Probanden während der Untersuchung und bei der Ergebnismitteilung;
- *Schweigepflicht* des Untersuchers gegenüber (unbefugten) Dritten;

- *Datenschutz* bei Aufbewahrung und Vernichtung von Informationen.

Zusammengefaßt finden sich ethisch-juristische Verpflichtungen beispielsweise

- in der „*Berufsordnung für Psychologen*“ (dpv, 1986),
- in den „*Leitsätzen zur Dokumentation klinisch-psychologischer/psychotherapeutischer Interventionen*“ (dpv, 1994 a),
- in den „*Richtlinien für die Erstellung Psychologischer Gutachten*“ (dpv, 1994 b),
- in den ethischen Grundsätzen, welche die APA 1992 veröffentlicht hat: „*Ethical Principles of Psychologists and Code of Conduct*“,
- in einschlägigen *Lehr- oder Handbüchern* (Amelang & Zielinski, 1994; Guthke, Böttcher & Sprunge, 1990; Jäger, R. S. & Petermann, 1995; Jessnitzer, 1980; Westhoff & Kluck, 1991; Wottawa & Hossiep, 1987; Zuschlag, 1992).

Aus zwei solcher Kompendien zitiert Kasten 12-1 einige Sätze des Vorwortes.

Kasten 12-1:

Aus dem Vorwort der BDP- und der APA-Normen

Aus dem Vorwort der BDP-Normen (dpv, 1986)

„Das berufliche Handeln des Psychologen, sei er nun als Arbeits- und Organisationspsychologe, als Klinischer Psychologe, in der Schul- und Pädagogischen Psychologie, Forensischen Psychologie oder in Lehre und Forschung tätig, ist geprägt von der besonderen Verantwortung, die der Psychologe gegenüber seinen Klienten/Patienten besitzt. Um helfen zu können, benötigt er ihr Vertrauen.

In weit höherem Ausmaß, als viele Berufsangehörige wissen, ist der Beruf des Psychologen in unsere Rechtsordnung integriert und von ihr abhängig.

Eine Berufsordnung ist stets auch Ausdruck des Selbstverständnisses eines Berufes. Sie vermittelt den Berufsangehörigen eine gültige Orientierung für ihre praktische Arbeit und setzt Maßstäbe, anhand derer psychologische Tätigkeiten öffentlich überprüfbar werden.“

Aus dem Vorwort der APA-Normen (APA, 1992)

„In the process of making decisions regarding their professional behavior, psychologists must consider this Ethics Code, in addition to applicable laws and psychology board regulations.

If the Ethics Code establishes a higher standard of conduct than is required by law, psychologists must meet the higher ethical standard.

If the Ethics Code standard appears to conflict with the requirements of law, then psychologists make known their commitment to the Ethics Code and take steps to resolve the conflict in a responsible manner.

If neither law nor the Ethics Code resolves an issue, psychologists should consider other professional materials and the dictates of their own conscience, as well as seek consultation with others within the field when this is practical.“

12.1 Zusammenfassung zu Kapitel 12

Ethische und juristische Ansprüche sind der diagnostischen Arbeit immanent, weil sie zu tun hat mit Selbst- und Fremdbestimmung des Probanden. Denn Diagnostik und Intervention berühren Sachverhalte, die einer Sphäre der Selbstbestimmung angehören.

Beispiele für Einzelpflichten:

- keine Demütigung, keine Nötigung während einer Untersuchung,
- Schutz der Privatsphäre des Probanden,
- Schweigepflicht des Untersuchers gegenüber (unbefugten) Dritten,
- Datenschutz.

12.2 Kontrollfragen zu Kapitel 12

- Diagnostik/Intervention und ihre Einbindung in einen ethisch-juristischen Kontext.
- Konkrete Beispiele für ethisch-juristische Pflichten.
- Beispiele für Kompendien, die ethisch-juristische Pflichten auflisten.

Zwei diagnostische Grundaufgaben: *Klassifikation und Selektion*

Psychodiagnostik ist eine Methodologie, die darauf abzielt, praktische Probleme (in Einzelfällen) zu lösen - aufgrund psychologischen Wissens und psychologischer ‚Techniken‘, wie sie in verschiedene Teildisziplinen bereitgestellt werden, beispielsweise in der Allgemeinen oder der Differentiellen Psychologie, in der Sozial- oder der Entwicklungspsychologie.

Die diagnostische Grundaufgabe, „praktische Probleme lösen zu helfen“, läßt sich aufgliedern in unterschiedliche Teilaufgaben. Zwei davon sind Klassifikation und Selektion (vgl. Cronbach, 1964, 17).

Eine solche Ausgliederung läßt sich verständlich machen aus dem ethisch-sozialen Kontext der Diagnostik. In diesem Rahmen stehen sich Individuum und Gesellschaft gegenüber und sind aufeinander bezogen. Beide ‚Partner‘ - der einzelne und die Gesellschaft - können einem Diagnostiker Aufgaben stellen:

1. *Ein einzelner kann ihn bitten, ihm zu helfen, in der ‚Bildungswelt‘ den ihm angemessenen ‚Ort‘ zu erkennen. Beispielsweise kann ein Zehnjähriger Hilfe suchen bei der Wahl der ‚richtigen‘ weiterführenden Schule. - In Fragestellungen dieser Art sehen wir die Aufgabe angelegt, die als Klassifikation bestimmt wird.*

Klassifikation bezeichnet einen diagnostischen Prozeß, in dem „einer Person bestimmte **sozial definierte** (sozial konstruierte) Dispositionen zugeschrieben werden. Ziel der Persönlichkeitsdiagnostik ist es dann, die Berechtigung dieser Zuschreibung zu überprüfen“ (Westmeyer, 1993, 508).

2. *Eine Gruppe der ‚Gesellschaft‘ kann einem Diagnostiker vorschlagen, ihr zu helfen, für eine vorgegebene ‚Stelle‘, die ‚geeignete‘ Person zu finden. Beispielsweise sei in einer Schule die Stelle des Rektors ausgeschrieben. Das verantwortliche Gremium suche Hilfe bei der Aufgabe, unter den zehn Bewerbern den ‚am besten geeigneten‘ Lehrer zu entdecken. - In Fragestellungen dieser Art sehen wir die Aufgabe angelegt, die als Selektion beschrieben wird.*

Selektion bezeichnet „einen gesellschaftlichen Sortiervorgang (Siebung). Die Auswahlkriterien sind kulturell bedingt und bestimmen damit die Art und Weise der sozialen Auslese. Aufgrund gemeinsamer Eigenschaften, Fä-

higkeiten oder Merkmale bilden sich Gruppen, die die Auslese bewußt oder unbewußt beeinflussen” (Humboldt-Psychologie-Lexikon, 1990, 44).

Wie zwischen einzelнем und Gesellschaft sowohl Spannung als auch Einvernehmen herrschen, so besteht *analog* zwischen Klassifikation und Selektion ein Verhältnis sowohl des Gegensatzes als auch der Ergänzung: dies sollen zwei Teilkapitel verdeutlichen:

- Klassifikation (13.1),
- Selektion (13.2).

Es folgen eine Zusammenfassung (13.3) und eine Reihe von Kontrollfragen (13.4).

13.1 Klassifikation

Bei der Klassifikation stellt sich die Aufgabe, für einen Probanden die ‚richtige(n)‘ Merkmalsklasse zu suchen.

Beispiel: *Ein Sechzehnjähriger bittet einen Psychologen um Hilfe bei der Berufswahl. Was ist der Gegenstand dieses Anliegens? In einer Vielzahl von Berufen den ‚angemessenen Beruf‘ zu entdecken!*

Allgemein läßt sich die diagnostische Aufgabe bei einer Klassifikation als Frage formulieren: In welche Merkmalsklasse gehört der Proband? Zu **suchen ist demnach die ‚Merkmalsklasse‘, die für einen Probanden ‚angemessen‘ ist.** Ein Schema in Kasten 13-1 kann dies veranschaulichen.

Kasten 13-1:
Klassifikationsaufgabe - schematisiert

Gegeben ist Proband 1	In welche Merkmalsklasse gehört Proband 1?	► In Klasse A? ► In Klasse B? ► In Klasse C?
-----------------------	--	--

Von Klassifikation spricht man also, wenn die diagnostische Bemühung vor allem darauf zielt, für eine Person die geeignete ‚Merkmalsklasse‘ zu finden. Gibt es mehrere Probanden, soll im Idealfall für jeden die ‚richtige‘ Zuordnung gefunden werden (vgl. Baumann, 1990; Janke, 1982; Klauer, 1987, 58-94).

Beispiele *für ‚Merkmalsklassen‘ denen sich ein Proband zuordnen läßt, sind Schularten, Stellen in einem Betrieb, Beschreibungsdimensionen eines Gutachtens.*

Voraussetzungen einer Klassifikation

Um die Aufgabe einer Klassifikation zu lösen, muß der Diagnostiker drei Festlegungen treffen:

- Er muß die Klassen genau **definieren**.
- Er muß **Kriterien** angeben, die bestimmte Leistungen abgrenzen (gegebenenfalls bestimmte Leistungsbereiche abstufen).
- Er muß **Entscheidungsregeln** formulieren, die besagen, bei welcher Leistung ein Proband einer Klasse zugewiesen wird.

Liegen quantifizierte Angaben vor, so lassen sich Werte angeben, welche Klassen (nach oben und nach unten) begrenzen. - Des weiteren lassen sich Werte (gegebenenfalls Intervalle) festlegen, die bestimmte Leistungen umschreiben. - Dann läßt sich vereinbaren, welche Werte jemand erreichen muß (welche Leistungen er somit zu erbringen hat), um einer bestimmten Klasse zugewiesen zu werden.

Als Beispiel seien Schulnoten genannt. Die Noten lassen sich als Merkmalsklassen auffassen, denen Schüler gemäß ihren Leistungen zuzuweisen sind. Eine Zuweisung ist valide,

- *wenn die Notenintervalle wohldefiniert,*
- *wenn als Kriterien die Leistungen präzise abgegrenzt und abgestuft sind,*
- *wenn die Regeln klar umschrieben sind, die festlegen, bei welcher Leistung ein Schüler welcher Notenklasse zugewiesen wird.*

Auf mathematische Klassifikationsmodelle, etwa Faktoren- oder Diskriminanzanalyse, sei hier nur verwiesen. Janke stellt sie ausführlich vor (1982).

HINWEIS: Werden Probanden **auf nur einer Dimension** gruppiert - etwa auf einem Test-Score - spricht man von **Plazierung**.

Richtige und falsche Zuordnung

Richtig ist eine Zuordnung, wenn ein ‚Merkmalsträger‘ jener Klasse zugewiesen wird, die das Merkmal repräsentiert, das er besitzt. - Typisch sind zwei Fälle richtiger Klassifikation:

- Der ‚Fähige‘ wird der Klasse der ‚Fähigen‘ zugewiesen, man spricht von „wahren Positiven“. **Beispiele:** *Begabte werden als begabt eingestuft, Gesunde als Gesunde.*
- Der ‚Unfähige‘ wird der Klasse der ‚Unfähigen‘ zugewiesen, man spricht von „wahren Negativen“. **Beispiele:** *Unbegabte werden als unbegabt eingestuft, Neurotiker als Neurotiker*

Bei Klassifikationen ist mit **Fehlern** zu rechnen. Falsch ist eine Zuordnung, wenn ein ‚Merkmalsträger‘ einer Klasse zugeordnet wird, die ein Merkmal repräsentiert, das er nicht besitzt. - Typisch sind zwei Arten von Fehlzuweisungen:

- Probanden, die *fähig* ‘sind, werden der Klasse der *„Unfähigen“* zugewiesen, man spricht von *„falschen Negativen“*. Es handelt sich um den sogenannten **a-Fehler**. - **Beispiele:** *Begabte werden als unbegabt eingestuft, Gesunde als Neurotiker.*
- Probanden, die *„unfähig“* sind, werden der Klasse der *„Fähigen“* zugeordnet, man spricht von *„falschen Positiven“*. Es handelt sich um den sogenannten **J-Fehler**. - **Beispiele:** *Unbegabte werden als begabt eingestuft, Neurotiker als gesund.*

Kasten 13-2 soll die beiden Arten von Klassifikationen (richtig/falsch) veranschaulichen (Amelang & Zielinski, 1994, 272; Klauer, 1987, 61).

Kasten 13-2:
Richtige und falsche Klassifikationen

		Ein Proband ist ,in Wirklichkeit‘	
		unfähig	fähig
Der Beurteiler stuft einen Probanden ein als	unfähig	<i>Richtige Negative</i> <i>Richtige Zuordnung:</i> Der Proband ist in Wirklichkeit unfähig, und der Beurteiler stuft ihn als unfähig ein.	<i>Falsche Negative</i> <i>a-Fehler</i> <i>Fehler:</i> Der Proband ist in Wirklichkeit <i>fähig</i> , aber der Beurteiler stuft ihn als unfähig ein.
	fähig	<i>Falsche Positive</i> <i>β-Fehler</i> <i>Fehler:</i> Der Proband ist in Wirklichkeit unfähig, aber der Beurteiler stuft ihn als <i>fähig</i> ein.	<i>Richtige Positive</i> <i>Richtige Zuordnung:</i> Der Proband ist in Wirklichkeit <i>fähig</i> , und der Beurteiler stuft ihn als <i>fähig</i> ein.

Welchen Fehler soll ein Beurteiler eher in Kauf nehmen: den a-Fehler (falsche Negative) oder den β-Fehler (falsche Positive)?

Er muß abwägen, welchen Nutzen und welche Kosten eine Entscheidung einschließt. Die Schwierigkeit einer solchen *Kosten-Nutzen-Schätzung* läßt sich artikulieren in Fragen wie den folgenden (Klauer, 1987, 62):

- Was ist ,schlimmer‘: *einen „Fähigen“ als „unfähig“ zu erklären* (a-Fehler) oder *einen „Unfähigen“ als „fähig“ einzustufen* (β-Fehler)?
- Wie soll ein solcher Fehler *gewichtet* werden?
- Welche *Größen* gehen in die Gewichtung ein?
- Lassen sich die Gewichte *quantifizieren* und *vergleichen*?
- Lassen sich *Fehler- Wahrscheinlichkeiten* angeben?

In den Prozessen solcher **Kosten-Nutzen-Schätzungen** dürften subjektive Gewichtungen unvermeidbar sein. Da die Fragen allgemeine Probleme der diagnostischen Situation betreffen, seien die Antwortversuche in einem eigenen Kapitel zusammengetragen (S. 373).

Bevor wir die Selektion besprechen, sei in Kasten 13-3 auf zwei spezielle Fälle der Klassifikation verwiesen.

Kasten 13-3:
Zwei spezielle Fälle von Klassifikation: Real- und Idealnormen

Die generelle Bedeutung von Klassifikation besagt ‚Zuweisung eines Probanden zu einer Merkmalsklasse‘. Zwei spezielle Fälle seien hervorgehoben:

1. Klassifikation nach sogenannten **Realnormen**:

Werden Personen klassifiziert nach Realnormen, so wird Bezug genommen auf Werte, die an Eich- oder Normstichproben gewonnen wurden. Es handelt sich um jene Art von Klassifikation, nach der die klassische Testtheorie vorgeht.

Realnormen liefern Vergleichswerte, die es ermöglichen, anzugeben, welche Position ein Proband einnimmt bezüglich der Werte einer Eich- oder Normstichprobe. Berechnet wird eine Kennzahl, die das Verhältnis der Leistung eines Einzelnen zu den Leistungen einer Stichprobe angibt (S. 111).

2. Klassifikation nach sogenannten **Idealnormen**:

Werden Personen klassifiziert nach Idealnormen, so wird Bezug genommen auf Kriterien, die ein „Ideal“ repräsentieren. Es ist jene Art von Klassifikation, nach der die kriteriumsorientierte Testtheorie vorgeht (S. 129).

Erwähnt seien zwei Varianten:

- Die **Medizin** orientiert sich an dem „Ideal“ einer „Freiheit von Krankheit“, besser noch: an dem Ideal der „Leiblichen Gesundheit“, wie schwierig es auch sein mag, Gesundheit zu „definieren“.
- Die **Psychotherapie** orientiert sich an dem „Ideal“ einer „Neurosenfreiheit“, besser noch: an einem Konzept der „Seelischen Gesundheit“, wiederum - wie schwierig es auch sein mag, dieses Konzept zu „definieren“.

13.2 Selektion

Bei dem anderen Ansatz diagnostischer Aufgaben, der Selektion, ist eine ‚Merkmalsklasse‘ vorgegeben, für die der ‚richtige‘ Proband zu suchen ist.

Beispiel: Ein Unternehmen A bietet Lehrstellen an, um die sich Sechzehnjährige bewerben sollen. Ein Unternehmen B schreibt einen Managerposten für ‚Führungskräfte‘ aus.

Hier läßt sich die diagnostische Aufgabe in eine Frage fassen, die komplementär zu der des ersten Ansatzes lautet: Welcher Proband besitzt die Merkmale, die ihn für die angebotene ‚Stelle‘ als geeignet ausweisen? Zu suchen ist demnach ein Proband, der einer definierten ‚Merkmalsklasse‘ am besten entspricht. Ein Schema in Kasten 13-4 kann dies veranschaulichen.

Kasten 13-4:
Selektionsaufgabe - schematisiert

Gegeben ist Stelle A	Welcher Proband ist für Stelle A am geeignetsten?	► Proband 1? ► Proband 2? ► Proband 3?
----------------------	---	--

Von Selektion spricht man also, wenn die diagnostische Bemühung vor allem darauf zielt, für eine definierte ‚Merkmalsklasse‘ (etwa eine Stelle) die geeignete Person zu finden. Selektion schließt die Möglichkeit der Ablehnung ein.

Beispiele vordefinierter ‚Merkmalsklassen‘, für die eine Person zu suchen ist, sind Stellen, die ein Betrieb anbietet, oder Studienplätze, die zur Verfügung stehen.

Aufgabe bei der Selektion ist es, die ‚Merkmalsklasse‘ genau zu beschreiben und an dieser Beschreibung (als Kriterium) die Bewerber zu ‚messen‘ (S. 343).

Basisrate, Selektionsrate, Validität

Drei Faktoren sollte der Diagnostiker bei der Selektion berücksichtigen:

- die Basisrate,
- die Selektionsrate und
- die Validität der eingesetzten Verfahren.

(Siehe etwa: Dorsch, 1994, 76, 86; Kompa, 1984,98-105; Lienert & Raatz, 1994, 389-393; Wottawa, 1980, 131-138!)

Die **Basisrate** (BR) bezeichnet die Häufigkeit, mit der ein bestimmtes Merkmal in einer bestimmten Gruppe auftritt. Geht man von einem konkreten Beispiel aus - etwa einer Stelle, um die sich Bewerber bemühen - dann definiert sich die Basisrate als Verhältnis der Zahl geeigneter Bewerber (N_c) zur Bewerbergesamtzahl (N_g): $BR = N_c/N_g$.

Dabei geht **es** um die **tatsächlich** Geeigneten, eine Zahl, die in der Regel nicht bekannt und kaum exakt bestimmbar ist. Das Problem besteht darin, die Zahl dieser geeigneten Personen zu schätzen.

Die **Selektionsrate** (SR) bezeichnet den relativen Anteil der auszulesenden Personen an der Gesamtzahl der Personen, die sich der Auslese stellen. Geht man wieder von dem Beispiel eines Stellenangebotes aus, dann definiert sich die Selektionsrate als Verhältnis offener Stellen (N_o) zur Bewerbergesamtzahl (N_g): $SR = N_o/N_g$.

Beispiel:

- 123 Kandidaten bewerben sich um 13 Sekretärsstellen. Es sei angenommen, daß unter den Probanden 25 Personen für die Stellen geeignet sind. Die Basisrate bestimmt sich dann wie folgt: $BR = N_c/N_g = 25/123 = 0.20$. Die **Basisrate beträgt 0.20**; das heißt, 20 Prozent der 125 Kandidaten sind ‚geeignete‘ Bewerber

- Die Selektionsrate bestimmt sich wie folgt: $SR = N_o/N_g = 13/123 \approx 0.11$. Die **Selektionsrate beträgt 0.11**; das heißt, nur etwa elf Prozent der Bewerber können eine Stelle bekommen.
- In dem Beispiel ist die Basisrate (0.20) größer als die Selektionsrate (0.11). Die Aufgabe besteht nun darin, aus der Zahl der 125 Kandidaten dreizehn der 25 ‚Geeigneten‘ zu identifizieren (günstigenfalls die dreizehn ‚Besten‘). Bei diesem Anliegen kann die Beachtung der Validität eines Verfahrens weiterführen.

Validität und Selektion

Was können valide Verfahren bei der Selektion leisten? **Taylor und Russe11** (1939) haben ein **Tafelwerk** vorgelegt, das es ermöglicht, den Zusammenhang zwischen Basis- und Selektionsrate sowie der Validität eines Verfahrens in Form einer **Trefferquote** zu schätzen. Kasten 13-5 gibt einen kurzen Auszug.

Was besagen die Taylor-Russell-Tafeln für unser Beispiel, nach dem sich 123 Probanden um 13 Sekretärsstellen bewerben? Gegeben waren eine Basisrate von $BR = 0.20$, eine Selektionsrate von $SR = 0.11$. **Verwendet werde ein Test mit der Validität $r_{tc} = 0.65$.**

Für diesen Fall geben die Tafeln den Zusammenhang an wie folgt:

Basisrate	= 0.20
Selektionsrate	= 0.11
Trefferquote	= 0.64

Was besagen die Angaben?

Die **Basisrate** beträgt **0.20**. Das heißt: Unter der Gesamtzahl der 123 Bewerber sind 20 Prozent ‚tatsächlich geeignet‘. Demnach kann man bei einer Selektion durch **Losverfahren** mit 20 Prozent Treffern rechnen.

- Setzt man bei einer Basisrate von 0.20 und einer Selektionsrate von 0.11 ein Instrument ein, das eine **Validität von 0.65** besitzt, dann ist nach den Taylor-Russell-Tafeln eine **Trefferquote von 0.64** zu erwarten.
- **Das bedeutet:** Gegenüber dem Losverfahren (also gegenüber einer Zufallsauswahl mit 20 Prozent Treffern) ergibt sich ein **Zugewinn von 44 Prozentpunkten**.

Ein anderes Beispiel findet sich in Kasten 13-5.

Kasten 13-5:
Auszug aus den Taylor-Russell-Tafeln

für die Basisrate von		BR = 0.60,			
die Selektionsraten von		SR = 0.90 bis SR = 0.10,			
Tests mit der Validität von		r_{tc} = 0.95 bis r_{tc} = 0.05.			
<i>Quelle: Lienert und Raatz (1994, 421)</i>					
	Basisrate: BR = .60				
	Selektionsrate (SR)				
<i>Validität</i>	.90	.70	.50	.30	.10
.95	.61	.84	.97	1.00	1.00
.85	.66	.80	.91	.97	1.00
.75	.66	.77	.86	.93	.99
.65	.65	.14	.82	.89	.96
.55	.64	.71	.78	.84	.92
.45	.64	.69	.74	.80	.87
.35	.63	.67	.71	.75	.82
.25	.62	.65	.68	.71	.76
.15	.61	.63	.65	.67	.70
.05	.60	.61	.62	.62	.63
<i>Beispiel:</i> Gegeben sind					
- eine Basisrate von		BR = .60,			
- eine Selektionsrate von		SR = .30,			
- ein Test mit der Validität von r_{tc}		= .75.			
<i>Dieser Fall ist in kursiv-fetter Schrift angezeigt.</i>					
Nach den Taylor-Russell-Tafeln ist eine Trefferquote von 0.93 zu erwarten. <i>Das bedeutet:</i>					
Gegenüber einer Zufalls-Trefferquote von 60 Prozent (BR = .60) ergibt sich ein Zugewinn					
von 33 Prozentpunkten.					

Drei Hinweise:

- Die Taylor-Russell-Tafeln sind ausgelegt für Gruppenuntersuchungen.
Generell gilt: „Zwei Voraussetzungen beschränken die Anwendbarkeit der Taylor-Russell-Tafeln. Zum einen wird angenommen, daß eine lineare Beziehung zwischen Prädiktor- und Kriteriumswerten besteht. Zum anderen wird vorausgesetzt, daß das neue Verfahren unabhängig von den anderen Prädiktorinstrumenten des bestehenden Auswahlsystems ist“ (Kompas, 1984, 104).
- *Speziell gilt:* In der Individualdiagnostik dürfte sich der ‚Zugewinn‘ (die inkrementelle Validität) kaum abschätzen lassen, den die Zufügung eines ‚weiteren‘ Verfahrens bringt. Beispielsweise dürfte es bei der Beurteilung von Bewerbern für eine Führungsposition schwierig sein abzuschätzen, um wieviel sich die Gesamtbeurteilung ‚verbessert‘, wenn - zu den schon angewandten Verfahren - ein noch so valider Test hinzukommt. Ins Spiel kommen hier idiosynkratische Aspekte (der Steile und des Bewerbers).

Verschränkung von Klassifikation und Selektion

Selektion und Klassifikation lassen sich nicht disjunkt trennen: Selektion schließt Klassifikation ein; denn sie erfordert eine exakte ‚Klassifikation‘ der ‚Stelle‘ und der Probanden. - Klassifikation schließt Selektion ein; denn - von der einzelnen Zuweisungsklasse her gesehen - findet unter den Bewerbern eine Auswahl statt. - Doch ist die Perspektive in beiden Fällen unterschiedlich.

13.3 Zusammenfassung zu Kapitel 13

Zwei wichtige diagnostische Aufgaben wurden besprochen. Vereinfacht gilt: Bei einer Klassifikation soll ein Proband einer ihm adäquaten Merkmalsklasse zugeordnet werden. - Bei einer Selektion soll für eine vorgegebene ‚Stelle‘ der geeignete Proband ausgewählt werden.

Bei Klassifikationen ist mit Fehlern zu rechnen. Um zwei Beispiele zu bringen: Probanden, die ‚fähig‘ sind, werden der Klasse der ‚Unfähigen‘ zugewiesen (α -Fehler). - Probanden, die ‚unfähig‘ sind, werden der Klasse der ‚Fähigen‘ zugeordnet (β -Fehler).

Für die Selektion können die Taylor-Russell-Tafeln gute Dienste leisten. Sie schätzen die sogenannte Trefferquote für Selektionsentscheidungen, wenn drei Größen bekannt sind:

1. die Basisrate (der Anteil geeigneter Bewerber an der Bewerbergesamtzahl),
2. die Selektionsrate (der Anteil der auszulesenden Personen an der Bewerbergesamtzahl),
3. die Validität des eingesetzten Verfahrens.

13.4 Kontrollfragen zu Kapitel 13

Definition von Klassifikation.

Definition von Selektion.

Definition, Kriterien und Entscheidungsregeln einer Klassifikation.

Klassifikationsfehler.

Basisrate, Selektionsrate.

Validität der Verfahren und ihr Einfluß auf die Selektion.

Funktion der Taylor-Russell-Tafeln.

Verschränkung von Klassifikation und Selektion.

14. Kapitel

Zwei Wege der Entscheidungsfindung: *Statistische und Klinische Urteilsbildung*

Strategien zur Erreichung diagnostischer Ziele werden vor allem unter zwei Titeln aufgeführt: unter dem der Statistischen oder dem der Klinischen Urteilsbildung (Amelang und Zielinski, 1994, 259-261; Heil, 1995, 39-42; Jäger, R. S., 1982, 302-308; Kompa, 1984, 77-82; Meehl, 1954; Wiggins, 1973).

Als wichtiger Unterschied gilt die Explizitheit der Regeln sowohl der Datenerhebung als auch der *Datenkombination*. Wir besprechen vier Themen:

- die Statistische Urteilsbildung (14.1),
- die Klinische Urteilsbildung (14.2),
- diagnostische Urteilsbildung und diagnostische Ziele (14.3),
- Vorrang des Statistischen Urteils (14.4).

Es folgen eine Zusammenfassung (14.5) und die Vorgabe einiger Kontrollfragen (14.6).

14.1 Statistische Urteilsbildung

Wenn Daten quantifiziert vorliegen (etwa Test- oder Fragebogenscores) und ihre Kombination auf einem ausformulierten Algorithmus beruht, dann spricht man von Statistischer Urteilsbildung.

Beispiel: Bei einer Bewerberauswahl stehen Daten zur Verfügung aus Tests und Fragebogen, aus Akten und Gesprächen: alle Daten sind quantifiziert. Kombiniert werden sie in der Gleichung einer multiplen Regression.

Voraussetzung für diese Art diagnostischer Urteilsbildung ist, daß Vergleichsdaten (Algorithmen, Normen, Validitätswerte) vorliegen: z. B. aus früheren Untersuchungen zu denselben Fragestellungen bei vergleichbaren Stichproben (Jäger, R. S., 1982; Leichner, 1979, 141-144; Schmidtchen, 1975, 12-14; Seitz, 1977, 30-41).

Statistische Urteilsbildung läßt sich verstehen als Versuch, Regeln, die an Gruppen gewonnen wurden, auf individuelle Fälle anzuwenden. Das Indivi-

duum wird unter dem Blick allgemeiner Gesetzmäßigkeiten betrachtet. - Kasten 14-1 bringt ein Beispiel.

Kasten 14-1:
Ein vereinfachtes Beispiel Statistischer Urteilsbildung, hier:
mittels multipler Regression
Daten fiktiv!

Vor einem Schulübertritt (Grundschule - Realschule) werden Tests durchgeführt, die erfassen sollen: ‚Sprachliche Intelligenz‘ (SI) und ‚Nichtsprachliche Intelligenz‘ (NI), ‚Rechnerisches Denken‘ (RD) und ‚Konzentrationsfähigkeit‘ (KO).

Aus früheren Untersuchungen liege eine multiple Regressionsgleichung vor, die es erlaubt, den Schulerfolg (Y') nach der sechsten Klasse vorherzusagen:

$$Y' = 1,31 + 0,23 \text{ SI} + 0,21 \text{ NI} + 0,32 \text{ RD} + 0,27 \text{ KO}$$

Die Angaben seien in Stanine gemacht. Ein Schüler erreiche in:

SI = 6 Stanine (Sprachliche Intelligenz),

NI = 4 Stanine (Nichtsprachliche Intelligenz),

RD = 8 Stanine (Rechnerisches Denken),

KO = 7 Stanine (Konzentrationsfähigkeit).

Einsetzen in die Gleichung und Berechnung ergeben als Vorhersage für ‚Schulerfolg nach der sechsten Klasse‘: $Y' = 7,98$. Da der Mittelwert der Stanine-Skala bei 5 und die Standardabweichung bei 2 liegt, erreicht der Schüler einen Wert oberhalb des Durchschnittsbereiches. Man müßte ihm hohe Erfolgsaussichten zusprechen.

14.2 Klinische Urteilsbildung

Wenn quantitative und qualitative Daten vorliegen (etwa Testscores, Zeugnisse, Verhaltensbeobachtungen) und wenn ihre Kombination auf dem Fachwissen, der Erfahrung, der Intuition des Diagnostikers beruht, ohne daß die Regeln des Urteilsganges mit allen Elementen explizit genannt werden, dann spricht man von **Klinischer Urteilsbildung**.

Beispiel: Bei einer Bewerberauswahl stehen Daten zur Verfügung aus Tests und Fragebogen, aus Akten und Gesprächen: ein Teil ist quantifiziert, ein Teil nicht. Der Diagnostiker kombiniert sie nach den ‚Regeln‘ seiner Berufserfahrung.

Indiziert ist Klinische Urteilsbildung dort, wo

- keine vorgängigen Untersuchungen vorliegen,
- keine Verknüpfungsregeln expliziter Natur gegeben sind.

Klinische Urteilsbildung läßt sich verstehen als Versuch, die individuelle Einmaligkeit zu erfassen (Fähigkeiten, Interessen, Motive, Konflikte, Einstellungen, Abwehrhaltungen). Das Individuum wird als Inbegriff einer einzigartigen Merkmalskonstellation betrachtet, die sich nur idiosynkratisch erfassen, nur partiell mit allgemeinen Gesetzmäßigkeiten beschreiben läßt. - Kasten 14-2 bringt ein Beispiel (vgl. S. 445).

Kasten 14-2:
Ein vereinfachtes Beispiel Klinischer Urteilsbildung, hier:
mittels psychologischer Begutachtung

In einer Familienrechtssache geht es um einen Streitfall zwischen Eheleuten, die getrennt leben, aber nicht geschieden sind: Herr L. will seinen zehnjährigen Sohn Tobias nach der Grundschule in ein Internat schicken, dem eine Realschule angegliedert ist. Er sieht dort günstigere Lernbedingungen. Die Mutter widerspricht dieser Absicht.

Das Familiengericht wird angerufen. Es beauftragt einen Psychologen mit einer Begutachtung, in der geklärt werden soll, ob „eine Trennung von der Mutter zu Beginn des neuen Schuljahres zu schwerwiegenden Komplikationen führt“.

Liegt für eine solche diagnostische Fragestellung bereits ein formulierter ‚Algorithmus‘ vor? Damit ist nicht zu rechnen! Eine Klinische Urteilsbildung dürfte für die Begutachtung angezeigt sein.

Das Gutachten stützt sich auf ein Studium der einschlägigen Akten, eine Unterredung mit dem Vater, eine Unterredung mit der Mutter, eine Untersuchung von Tobias, in die eine Vielfalt von Verfahren einbezogen wurde.

Die Auswahl des Instrumentariums, die Auswertung der Verfahren (Gespräche, Verhaltensbeobachtung, Leistungs-, Persönlichkeitstests) sowie die Integration der Informationen zu einer Gesamtstellungnahme beruhen auf der pädagogischen und psychologischen Fachkenntnis des Diagnostikers.

Wie bei der Statistischen Urteilsbildung sollten die Argumente, ihre Gewichtung und Verknüpfung so transparent gemacht werden, daß die Stellungnahme den Adressaten durch ihre Stringenz überzeugt, allerdings nicht in Gestalt eines Algorithmus, sondern in Form eines schlüssigen fortlaufenden Textes.

14.3 Diagnostische Urteilsbildung und diagnostische Ziele

Statistische wie Klinische Urteilsbildung kann unterschiedlichen Zielen dienen, beispielsweise der Selektion und der Klassifikation.

Zwei **Beispiele sollen** diese Aussage erläutern:

- Eine **Selektion** von Führungskräften kann quantitative Daten (etwa aus Persönlichkeitstests) ebenso einschließen wie qualitative Angaben (etwa aus Gesprächen). Die Entscheidung ‚Bewerber angenommen‘ (Bewerber nicht angenommen) muß nicht auf einem expliziten Algorithmus beruhen.
- Eine **Klassifikation**, etwa die Zuweisung eines Schülers zu einer Sonderschule, kann sich auf psychometrische Verfahren (etwa auf Schultests) und einen wohldefinierten Algorithmus stützen.

Welche Entscheidungsstrategie gewählt wird, hängt ab von dem Kenntnisstand, der zu einem Problem vorliegt, ebenso von den methodischen und personalen Kapazitäten, die dem Untersucher zur Verfügung stehen.

Für eine terminale Entscheidung können beide Urteilsmodelle unterschiedliche Schritte vorsehen:

- **Entweder** muß der Proband in einer festgelegten Zahl von Merkmalen **bestimmte Marken** erreichen (also bestimmte Ausprägungsgrade der Merkmale). Ein noch so hoher Wert auf der Dimension A kann einen Wert auf der Dimension B **nicht kompensieren**, der unterhalb des markierten Ausprägungsgrades liegt (Cut-Off-Modell). - **Beispiel:** Bei einem Piloten-Anwärter vermag eine noch so vorzügliche Körperbeherrschung eine Farbenblindheit nicht auszugleichen.
- **Oder** der Proband muß in festgelegten Dimensionen eine **festgelegte Gesamtmarke** erreichen. Die einzelnen Dimensionen werden gewichtet. Hochleistungen in einer Dimension können Schwächen in anderen Dimensionen ausgleichen (statistisches Modell: Multiple Regression). - **Beispiel:** In einer Schullaufbahn kann hohe Leistungsmotivation das Defizit in einer einzelnen kognitiven Teilfunktion ausgleichen.

14.4 Vorrang des Statistischen Urteils?

Statistisches wie Klinisches Urteil lassen sich auf vielfältige Weise realisieren (Jäger, 1982). In der Forschung und bei Reihenuntersuchungen dürfte die Statistische, in der Individualdiagnostik dürfte die Klinische Urteilsbildung überwiegen.

Berechtigt scheint auch die Forderung: *Immer dann, wenn für ein Problem Algorithmen formuliert und alle Angaben quantifiziert vorliegen, sollte der Diagnostiker die Statistische Urteilsbildung bevorzugen!*

Nur dürfte er für die meisten Probleme weder angemessene Algorithmen noch vollständig quantifizierte Daten vorfinden.

Zuweilen wird behauptet, das Statistische Urteil sei dem Klinischen generell überlegen (Kompas, 1984, 78-85; Leichner, 1979, 142-144). Verwiesen wird auf die Analysen von Meehl (1954) und Sawyer (1968). Doch dürfte dieser Vorrang nicht so eindeutig begründet zu sein, wie behauptet wird. Denn die Untersuchungen, die Meehl und Sawyer gesammelt haben, sind zu disparat, die „Erfolgskriterien“ zu unscharf definiert, als daß sie ein eindeutiges Urteil erlaubten.

Eher dürfte zutreffen, was Jäger, R. S. sagt (1982, 308): „Unseres Erachtens muß Klinische und Statistische Urteilsbildung jeweils als eine Möglichkeit des Zugangs im Erkenntnisprozeß gesehen werden, wobei aber das Ziel darin besteht, das *Procedere* explizit zu machen.“

Bei komplexen Fragestellungen dürfte es sich empfehlen, beide Vorgehensweisen zu verbinden (etwa den einen Teilbereich durch Statistische, den anderen durch Klinische Urteilsbildung abzuklären).

14.5 Zusammenfassung zu Kapitel 14

Das Kapitel bespricht zwei Modelle diagnostischer Urteilsbildung: Bei Statistischer Urteilsbildung werden Daten erhoben und kombiniert nach einer vollständig quantifizierten (psychometrischen) Prozedur.

Bei Klinischer Urteilsbildung werden Daten erhoben und kombiniert nach Regeln, die auf der Erfahrung und Intuition des Diagnostikers beruhen und nicht vollständig quantifiziert sind.

Beide Urteilsmodelle können den Aufgaben der Klassifikation wie denen der Selektion dienen. Aber die beiden Urteilsmodelle können einander nicht vertreten:

- Wenn für ein Problem Algorithmen formuliert sind und die diagnostischen Angaben quantifiziert vorliegen, sollte das Statistische Urteil Vorrang erhalten.
- In vielen Fällen trifft weder das eine noch das andere zu: weder liegen entsprechende Algorithmen vor, noch sind alle Daten quantifiziert. In vielen diagnostischen Aufgaben ist darum das Klinische Urteil das angemessene Modell.

14.6 Kontrollfragen zu Kapitel 14

- Statistisches Urteil: Definition.
- Klinisches Urteil: Definition.
- Beispiele für Statistische/Klinische Urteile.
- Anwendungsfälle für das Statistische Urteil.
- Anwendungsfälle für das Klinische Urteil.
- Vorrang des statistischen Urteils?

15. Kapitel

Drei Ansätze für Diagnostik und Intervention

Ein Psychologe kann seine diagnostischen und interventiven Untersuchungen von unterschiedlichen Ansätzen her konzipieren.

Welchen Untersuchungsplan er entwickelt, sollte die Fragestellung bestimmen; doch spielen auch Dauer und Kosten einer Untersuchung eine Rolle.

Wir besprechen drei Beispiele. Eine Untersuchung bezieht sich auf

- Verhaltensperformanz oder Verhaltensdeskription (15.1),
- synchrone oder diachrone Verhaltensbetrachtung (15.2),
- Verhaltensstatus oder Verhaltensprozeß (15.3).

Es folgen eine Zusammenfassung (15.4) und eine Reihe von Kontrollfragen (15.5).

15.1 Verhaltensperformanz oder Verhaltensdeskription

Der Diagnostiker kann eine Untersuchung planen auf der Ebene der Performanz oder der Deskription.

Auf der Ebene der Performanz wird ein Merkmal dann gemessen, wenn ein Verfahren jene Verhaltensanteile *evoziert*, die das Zielmerkmal charakterisieren. Diesem Ansatz entspricht ein Untersuchungsplan, der von dem Probanden verlangt, daß er *Handlungen ausführt*, die das angezielte Merkmal kennzeichnen.

Beispiel: Eine Organisation schreibt eine Führungsposition aus. Die Bewerber müssen ein Verfahren durchlaufen, das ihnen auferlegt, alle Anforderungen zu **realisieren**, welche die Zielposition verlangt. Ein solches Verfahren ist beispielsweise das Assessment-Center (vgl. Kap. 23, S. 491).

Auf der Ebene der **Deskription** wird ein Merkmal dann erfaßt, wenn ein Verfahren jene Verhaltensanteile *beschreibt*, die das Zielmerkmal charakterisieren. Diesem Ansatz entspricht ein Untersuchungsverlauf, der von dem Probanden verlangt, daß er seine *kognitive Repräsentanz* über jene Verhaltensanteile *bekannt gibt*, die bei ihm erfaßt werden sollen.

Beispiel: Eine Organisation schreibt eine Führungsposition aus. Die Bewerber müssen an einem Einstellungsinterview teilnehmen, einem Verfahren, in dem sie sich selber schildern und präsentieren können.

Sowohl Performanz als auch Deskription haben im diagnostischen Prozeß ihre Bedeutung.

- Wo immer möglich, sollte der Untersucher Verfahren auswählen, die es erlauben, Verhalten auf der Ebene der Performanz zu erfassen.
- Doch gibt es Dimensionen, die für den Untersucher hochbedeutsam sind, sich einer Realisierung in der Untersuchung jedoch entziehen. **Beispiele** sind Motivstrukturen oder Zukunftspläne, kritische Lebensereignisse oder die Entwicklung von Partnerschaften. In Fällen wie diesen sind deskriptive Verfahren angezeigt.

Darüber hinaus ist festzuhalten: Daten, gewonnen mit deskriptiven Methoden, haben sich als valide Prädiktoren bewährt (Eckardt, 1977, 550; Esser, 1995, 654; Fisseni, Olbrich, Halsig, Mailahn & Ittner, 1993; Funke, U. & Schuler, 1986, 34; Mischel, 1981, 303).

Instrumente der Untersuchung

Der Performanz sind Verfahren zugeordnet wie Experimente, Leistungstests oder ‚Arbeitsproben‘, schließlich auch Einzelgespräche mit dem Ziel, eine Stichprobe verbalen **Verhaltens** zu ziehen.

Der Deskription sind alle Verfahren zugeordnet, die geeignet sind, bei dem Probanden die kognitive Repräsentanz über das Zielmerkmal abzurufen: Biographische Fragebögen ebenso wie Persönlichkeitsinventare oder unterschiedliche Formen des Interviews.

15.2 Synchrone oder diachrone Verhaltensbetrachtung

Aktueller oder biographischer Ansatz

Eine diagnostische Untersuchung kann auf zweierlei abheben:

- Sie kann darauf zielen, die aktuelle Situation eines Probanden zu ermitteln: dann handelt es sich um eine synchrone Betrachtung.
- Die Untersuchung kann aber auch darauf zielen, die biographische Entwicklung eines Individuums zu beschreiben: dann handelt es sich um eine diachrone Betrachtung.

Bei bestimmten Fragestellungen ist es angemessen, den **aktuellen Status** eines Probanden festzustellen. Es soll ein Bild seiner aktuell verfügbaren Fähigkeiten und Fertigkeiten gewonnen werden. Vor allem Eignungsuntersuchungen können sich auf dieses Ziel ausrichten.

Beispiel: Bewerber um medizinische Studienplätze bearbeiten den ‚Test für medizinische Studiengänge‘ (TMS: Trost et al., 1995). Dabei interessiert die Untersucher was die Bewerber in der aktuellen Prüfungssituation leisten. Das aktuelle Ergebnis (der Testscore) dient als **ein** Schätzwert der Studieneignung.

Bei anderen Fragen ist es ‚richtiger‘, die aktuellen Fähigkeiten einzubetten in ihre Entstehungsgeschichte. Gegenstand der Untersuchung ist dann die **biographische Entwicklung** eines Probanden. Dieses Vorgehen empfiehlt sich, wenn es um Sachverhalte wie Partnerschaftskonflikte oder neurotische Störungen geht, aber auch bei bestimmten Eignungsfragen, etwa bei der Berufswahl (Fuchs, W., 1982).

Beispiel: Bewerber um medizinische Studienplätze werden über ihren bisherigen Studienweg exploriert. Angenommen wird dabei, daß unterschiedliche Personen gelernt haben, ihre Fähigkeiten in unterschiedlicher Weise einzusetzen. So können verschiedene Probanden, die gleich hohe Intelligenz besitzen, ihre Prüfung nach Strategien vorbereiten, die erheblich divergieren, aber gleich erfolgreich sind (Fisseni, Olbrich, Halsig, Mailahn & Ittner, 1993).

Instrumente der Untersuchung

Keine diagnostische Methode dürfte sich allein für den biographischen oder allein für den aktuellen Ansatz eignen. Doch ergeben sich unterschiedliche Schwerpunkte.

Bei einer Orientierung an dem ‚aktuellen Status‘ leisten nützliche Dienste

- *Tests und Persönlichkeitsinventare*, soweit sie relativ stabile Verhaltensmuster erfassen,
- *Explorationen*, soweit sie die gegenwärtige Situation verständlich machen,
- *projektive Verfahren*, soweit sie den Diagnostiker auf aktuelle Techniken und Strategien aufmerksam machen.

Bei einer Orientierung an der biographischen Entwicklung empfehlen sich

- *Anamnese und Exploration*, soweit sie Verlauf und ‚kritische Lebensereignisse‘ erschließen,
- *projektive Verfahren*, soweit sie auf die Entstehung von Techniken und Strategien aufmerksam machen,
- (*Tests und*) *Biographische Fragebögen*, soweit sie gegenwärtige Verhaltensmuster auf dem Hintergrund ihrer Entstehung erhellen.

Verbindung von aktuellem und biographischem Ansatz

Aktueller und biographischer Ansatz sind nicht als Gegensätze zu betrachten, sondern als unterschiedliche diagnostische Zugänge, die abhängen vom Untersuchungsziel. Doch dürfte gelten: *Je ‚wichtiger‘ (je existentieller) eine diagnostische Frage, desto angemessener ist ein biographischer Zugang.*

15.3 Verhaltensstatus oder Verhaltensprozeß

Eng verwandt mit einer Erfassung von Daten aus diachroner oder synchroner Sicht ist ein Problem, das unter Titeln wie ‚Status-‘ oder ‚Prozeßdiagnostik‘ behandelt wird (Goldfried & Kent, 1976; Jäger, R. S., 1986, 1992; Jäger, R. S. & Scheurer, 1992; Pawlik, 1976; Schulte, 1976).

Statusdiagnostik wird gekennzeichnet „als eine Art der Diagnostik, bei der das Ziel im Vordergrund steht, einen psychologischen Ist-Zustand festzuhalten und die Feststellung diagnostisch oder prognostisch zu nutzen. Bei einer statusbezogenen Diagnostik wird von zeit-, situations- und populationsstabilen Merkmalen ausgegangen“ (Jäger, R. S., 1986, 89).

Beispiel: *Ein Behinderter beantrage den Führerschein. Er werde zu einer psychologischen Untersuchung verpflichtet. Zu entscheiden ist dann vorrangig, ob er über die kognitiven, motorischen, motivationalen Fähigkeiten verfügt, die erforderlich sind, ein Fahrzeug zu steuern.*

Prozeßdiagnostik wird gekennzeichnet durch einen Bezug, „bei dem mit Hilfe diagnostischer Methoden Veränderungen festgestellt werden können... Solche Veränderungen im psychologischen Kontext beziehen sich auf Verhaltens- und Erlebnisweisen, von denen angenommen wird, daß sie z.B. durch modifikatorische Maßnahmen beeinflusst worden sind“ (Jäger, R. S., 1986, 89).

Beispiel: *Ein Fernfahrer habe den Führerschein verloren „wegen Alkohols am Steuer“. Bei seinem Antrag, den Führerschein wiederzuerhalten, werde er zu einer medizinisch-psychologischen Untersuchung verpflichtet. Zu entscheiden ist dann nicht nur ob er über die kognitiven, motorischen, motivationalen Fähigkeiten verfügt, die erforderlich sind, ein Fahrzeug zu steuern. Vorrangig ist vielmehr zu entscheiden, ob er bereit ist, sein Verhalten in einem solchen Grade zu verändern, daß mit genügender Wahrscheinlichkeit die Prognose gestellt werden kann, in Zukunft werde sein Fahrverhalten den Regeln des Straßenverkehrs entsprechen.*

Die ‚Definitionen‘ lassen erkennen: Statusdiagnostik ist eher einer ‚Eignungsdiagnostik‘ affin, Prozeßdiagnostik dient eher der Vorbereitung von Interventionen.

Doch dürfte der Unterschied nicht essentieller Natur sein:

- **Statusdiagnostik kann übergehen in Prozeßdiagnostik**, Ein Diagnostiker habe den ‚Zustand‘ eines Probanden erfaßt. Dann kann er das Ergebnis mit ihm besprechen: *Er kann Zusammenhänge aufhellen, Schwierigkeiten erklären und kann so die Problemsicht des Probanden verändern.* Denkt man diesen Prozeß weiter und interpretiert ihn im Sinne einer ‚interaktiven Diagnostik‘, dann läßt sich ein Untersuchungsverlauf als eine Prozedur betrachten, die einer pädagogischen oder therapeutischen Intervention sehr nahekommt.

- **Prozeßdiagnostik setzt Statusdiagnostik voraus.** Veränderungen werden gemessen, indem zu unterschiedlichen Zeitpunkten ein Status erhoben und die Sequenz der ‚Zustände‘ verglichen wird. *Erhoben werden somit querschnittliche Statusbefunde. Miteinander verglichen, ergeben sie das Bild eines längsschnittlichen Verlaufs.*

Dies trifft auch dann zu, wenn ein Proband unmittelbar über Veränderungen Auskunft gibt, beispielsweise wenn er gefragt wird: Hat sich zwischen erster und vierter Therapiestunde etwas verändert? In diesem Falle ist es der Klient und nicht der Untersucher, der zwei Zustände vergleicht und ihren Unterschied als Änderung interpretiert.

Instrumente der Untersuchung

Was das diagnostische Instrumentar angeht, so dürften sich Ähnlichkeiten ergeben zu einer diachronen oder synchronen Verhaltensbetrachtung. Keine Methode dürfte nur für die Status- und keine nur für die Prozeßdiagnostik verwendbar sein. - Die folgenden „Hinweise zur Veränderungsmessung“ benennen verschiedene Instrumente.

Hinweise zur Veränderungsmessung

Prozeßdiagnostik erfordert Veränderungsmessung: Der Zustand einer Person wird zu wenigstens zwei Zeitpunkten erfaßt. Zwischen den zwei Meßzeitpunkten, so wird angenommen, verändert sich das Verhalten aufgrund einer Intervention.

Das Problem sei skizziert in Anlehnung an Petermann (1989) sowie an Jäger, R. S. und Scheurer (1992).

Voraussetzungen: Exakt erfaßt werden können Veränderungen dann, wenn

1. parallele *Situationen* t_1 bis t_k konstruiert werden und
2. parallele *Messungen* m_1 bis m_k diese Situationen erfassen.

Zu 1.: Wann sind - gefragt aus Sicht der klassischen Testtheorie - die Situationen t_1 bis t_k streng parallel? Wenn in allen Situationen die *wahren Werte* vollständig äquivalent sind (oder nur um Zufallsbeträge variieren)!

Zu 2.: Wann sind die Messungen m_1 bis m_k streng parallel? Wenn ihre Kennwerte exakt gleich sind (oder nur um Zufallsbeträge variieren)!

Schwierigkeiten: Bei der Konstruktion paralleler Situationen und paralleler Messungen (oder Meßinstrumente) ergeben sich erhebliche Schwierigkeiten. Vielfältige Einflüsse vermindern oder verhindern die Parallelität: Transfer von Erfahrungen aus der einen Situation auf die andere, affektive und motivationale Reagibilität (etwa Abneigung oder Übersättigung), veränderte (soziale) Mitwelten und (sachliche) Umwelten.

Lösungsvorschläge

Bisher hat kein Lösungsvorschlag alle Schwierigkeiten behoben,

HINWEIS: Die Vorschläge a bis c bleiben im Rahmen der klassischen Testtheorie, die Vorschläge d bis f greifen über sie hinaus.

a) *Experimentalgruppe versus Kontrollgruppe:* Es handelt sich um ein klassisches Design. Vorgesehen sind Experimental- und Kontrollgruppe, Vor- und Nachtest. Nur die Experimentalgruppe durchläuft eine Intervention. Im Nachtest sollten/müßten Experimental- und Kontrollgruppe sich unterscheiden.

Probleme:

- Sind die Situationen in Vor- und Nachtest exakt parallel?
- Läßt sich das Design auf den Einzelfall übertragen?

b) *Paralleltests:* Die Probanden werden zu wenigstens zwei Zeitpunkten t_1 mit dem Test A, zum Zeitpunkt t_2 mit dem Paralleltest B gemessen.

Probleme:

- Sind Test A und Test B streng parallel?
- Bleiben Lerneffekte, modifizierende Umwelteinflüsse, veränderte Stimmungslagen ausgeschlossen?
- Gibt es genügend Paralleltests, wenn mehr als zwei Messungen vorgesehen sind?

c) *Einsatz änderungssensitiver Instrumente:* Betroffene Personen dienen selber als Meßinstrument. Sie beschreiben die Änderungen, die sie wahrnehmen (perceived changes of behavior).

Zwei Varianten:

Erstens, der ‚behandelte‘ Proband zeigt die Änderung an. Beispielfrage: „Bin ich innerlich ruhiger geworden?“ Drei Antwortalternativen: (1) „Keine Änderung!“ - (2) „Ja, ich bin ruhiger geworden.“ - (3) „Nein, ich bin sogar unruhiger geworden.“

Zweitens, eine Drittperson schätzt die Änderung ein, nach dem gleichen Schema - nur in dritter Person.

Probleme:

- Ist anzunehmen, daß die Probanden ihre Einschätzung auf streng parallele Situationen beziehen?
- Sind die Schätzungen verschiedener Probanden vergleichbar?

d) *Probabilistische Verfahren:* Ein Itemsatz A erfaßt den Zustand zum Zeitpunkt t_1 , ein Itemsatz B den Zustand zum Zeitpunkt t_2 , ein Itemsatz C den Zustand zum Zeitpunkt t_3 . Differenzen gelten als Indikatoren von Änderungen.

Probleme:

- Die Antwort auf **ein** Item darf nicht abhängen von der Antwort auf ein **anderes** Item (lokale stochastische Unabhängigkeit): Wie läßt diese Unabhängigkeit sich gewährleisten?
- Eine solche Unabhängigkeit vorausgesetzt: wäre sie für eine pädagogische und eine therapeutische Intervention überhaupt zu wünschen?

e) *Verwendung multivariater linearer Modelle:* Verwandt werden Varianz- und Faktorenanalysen.

- Bei *Varianzanalysen* können „Sprünge“ in der Sequenz von Meßwiederholungen als Interventionseffekte gedeutet werden.
- Bei der *Faktorenanalyse* können bei derselben Person mehrere Merkmale zu verschiedenen Zeitpunkten untersucht werden (sogenannte O-Technik). Erkennbar wird gegebenenfalls die gemeinsame Fluktuation verschiedener Merkmale zu verschiedenen Meßzeitpunkten.

Probleme:

- Sind die gemessenen Variablen voneinander unabhängig?
- Sind die Variablen nur linear verknüpft?
- Könnten sie nicht auch „multiplikativ aufeinander einwirken“?

f) *Zeitreihenanalysen:* Der Prozeßverlauf des Erlebens und Verhaltens wird in drei Komponenten zerlegt. *Trend* bezeichnet die Richtung des Verlaufes entlang der Zeitachse, *Oszillation* die Schwankungen „nach oben oder unten“ um den Trend-Mittelwert, *der Fehler* den unaufgeklärten Rest der Varianz.

Probleme:

- Beeinflussen in der Zeitreihe frühere Messungen nicht die späteren (serielle Abhängigkeit)?
- Schwanken die Werte nur um den Mittelwert, fluktuiert der Mittelwert nicht auch selber?

Resümee zu den Lösungsvorschlägen: „Zu Recht wird man . . . einwenden, daß derzeit der methodische und rechentechnische Aufwand in keiner entsprechenden Relation zum Nutzen für den praktisch arbeitenden Diagnostiker steht, zumal viele der angedeuteten Verfahren . . . nur unter Nutzung von großen Rechenanlagen eingesetzt werden können“ (Jäger, R. S. & Scheurer, 1995, 207).

Drei Aporien der Veränderungsmessung

Wenn Veränderungen gemessen werden sollen, ergeben sich aus der Syntax der klassischen Testtheorie drei Aporien:

- das Bedeutsamkeitsproblem (A),
- die Regression zur Mitte (B),
- das Reliabilitäts-Validitäts-Dilemma (C).

(A) Bedeutsamkeitsproblem

Zwei Personen spenden hundert Mark für ein Hilfswerk: Muß dann die gleiche Summe in den beiden Fällen auch das gleiche „bedeuten“? Sei *der eine* Spender der Leiter eines Metallunternehmens, *der andere* ein Metalldreher in diesem Betrieb. Keineswegs „bedeuten“ die hundert Mark für den Firmenleiter und seinen Dreher das gleiche. Die Relation des Spendenbetrages zum Gesamteinkommen der beiden bestimmt die Bedeutung mit.

Dieses Beispiel soll ein spezielles Problem einführen, das bei der Veränderungsmessung auftritt.

Nach der Syntax der klassischen Testtheorie zeigen gleiche Abstände zwischen Skalenwerten an, daß auch zwischen den Ausprägungen der Merkmale gleiche Abstände bestehen.

Zwei Beispiele:

1. Ein IQ von 115 liegt um fünf Einheiten über dem IQ von 110: Also ist die Intelligenz eines Probanden mit dem IQ 115 um fünf Ausprägungsgrade „höher“ als die Intelligenz eines Probanden mit dem IQ 110.
2. Ein IQ von 75 liegt um fünf Einheiten unter dem IQ von 80: Also ist die Intelligenz eines Probanden mit dem IQ 75 um fünf Ausprägungsgrade „geringer“ als die Intelligenz eines Probanden mit dem IQ 80.

Angesichts dieser Interpretation stellt sich die Frage: *Wie weit entsprechen sich psychometrische Meßwerte und ihre psychologische Bedeutung?*

Die Frage enthüllt ihre Brisanz, wenn sie erneut an Beispielen veranschaulicht wird.

- Zwei Läufer verkürzen ihre Laufzeit um den gleichen Betrag: Bei einem zweiten Lauf legen sie die Strecke von 100 Metern um 0.2 Sekunden schneller zurück als beim ersten Lauf. *Indessen* - der eine Läufer verbessert sich von 10 Sekunden auf 9.8, der andere von 13.2 auf 13 Sekunden. *Bedeutet der „physikalisch gleiche Zugewinn“ auch psychologisch das gleiche für die beiden Läufer?*
- Bei einer Raucher-Entwöhnung reduziert *eine* Rauchergruppe ihren Zigarettenkonsum von 50 auf 40, eine *andere* Gruppe von 10 auf Null. *Hat die Differenz in beiden Gruppen die gleiche Bedeutsamkeit?*

Resümee: Ein Interpret sollte Verhaltensänderungen nicht deuten allein nach dem Betrag psychometrischer Differenzen, sondern auch nach ihrer Bedeutsamkeit für die betroffenen Probanden.

(B) Regression zur Mitte

Bei Veränderungswerten gilt nach der Syntax der klassischen Testtheorie:

- Je *höher* ein Wert bei der Erstmessung liegt, desto wahrscheinlicher ist es, daß der Wert bei der Zweitmessung *absinkt* - er nähert sich dem Mittelwert.
- Je *niedriger* der Wert bei der Erstmessung liegt, desto wahrscheinlicher ist es, daß der Wert bei der Zweitmessung *anstiegt* - auch er nähert sich dem Mittelwert.

Verständlich wird dieser Effekt aus der Regressionsrechnung, wie Kasten 15-1 an einem Beispiel erläutert.

Anwendung: An einer Psychotherapie nehmen Klienten teil, die einen extrem hohen Neurotizismus-Score aufweisen. Wird nach einer Therapie der Neurotizismus erneut gemessen, so ist es wahrscheinlich, daß der zweite Meßwert gefallen ist und sich dem Mittelwert genähert hat. - *Zwei Interpretationen* bieten sich an:

1. Die therapeutische Intervention war effektiv, sie hat den Neurotizismus ‚reduziert‘.
2. Oder aber der „Rückschritt zur Mitte“ ist bloß ein statistischer Artefakt.

Kasten 15-1:

Regressionseffekt bei Veränderungsmessung - Veranschaulichung

Aus dem Wert eines Prätests A wird der Wert des Posttests B geschätzt. Die entsprechende Regressionsgleichung lautet:

$$B' = r_{AB} \frac{S_A}{S_B} (B_i - \bar{B}) + \bar{A}$$

Es bedeuten:

A : Werte im Prätest,

B : Werte im Posttest,

B' : durch Regression geschätzter Wert im Posttest,

r_{AB} : Korrelation des Prätests A mit dem Posttest B,

S_A, S_B : Standardabweichung von A, von B.

Zwei Fälle sollen den Regressionseffekt veranschaulichen!

Für die Werte von A und B wird gleiche Verteilung angenommen.

Es gelte: $S_A = S_B = 10$, $\bar{A} = \bar{B} = 100$. Sei $r_{AB} = 0.78$.

1. Liege der Wert A_i im Prätest hoch, hier bei 110, so sinkt der geschätzte Wert B' auf 107.8. - *Einsetzen:*

$$B' = 0.78 \frac{10}{10} (110 - 100) + 100 = 107.8$$

2. Liege der Wert A_i im Prätest niedrig, hier bei 90, so steigt der geschätzte Wert B auf 92.2. - *Einsetzen.*

$$B' = 0.78 \frac{10}{10} (90 - 100) + 100 = 92.2$$

Kommentar:

In beiden Fällen nähert sich der geschätzte Wert B dem Mittelwert 100. Darum die Bezeichnung „Regression zur Mitte“!

(C)
Reliabilitäts-Validitäts-Dilemma:

Die Axiome der klassischen Testtheorie setzen die Stabilität der gemessenen Merkmale voraus. Werden Merkmale wiederholt gemessen, etwa in einem Prä-Post-Design, und verändern sich die Scores, dann werden die Veränderungen „system-immanent“ als Fehler interpretiert.

Daraus ergeben sich paradoxe Konsequenzen:

1. Korrelieren Prätest-Score und Posttest-Score hoch ($r_{\text{prä, post}} \rightarrow 1$), dann ist die Messung ihrer Differenzen höchst unreliabel ($r_{\text{tt diff}} \Rightarrow 0$).
2. Hochreliabel ist die Messung der Differenzen nur ($r_{\text{tt diff}} \rightarrow 1$), wenn Prätest-Score und Posttest-Score niedrig miteinander korrelieren ($r_{\text{prä, post}} \Rightarrow 0$).

Kommentar:

- Im Fall 1 zeigt die hohe Korrelation zwischen Vor- und Nachtest einen erwünschten Effekt an ($r_{\text{prä, post}} \Rightarrow 1$): *Die Merkmale in Vor- und Nachtest erweisen sich als gleich.* Indessen - ihre Messung ist diagnostisch wertlos, weil völlig fehlerhaft ($r_{\text{tt diff}} \Rightarrow 0$).
- Im Fall 2 zeigt die hohe Reliabilität ebenfalls einen erwünschten Effekt an ($r_{\text{tt diff}} \rightarrow 1$): *Die Messung der Merkmale ist zuverlässig.* Indessen - es bleibt völlig ungeklärt, was gemessen wird. Es fragt sich, ob Prätest und Posttest dasselbe gleiche Merkmal messen ($r_{\text{prä, post}} \Rightarrow 0$).

Anwendung: An einer Psychotherapie nehmen Klienten teil, die bei einer ersten Messung einen hohen Neurotizismus-Score aufweisen. Bei einer zweiten Messung sei der Neurotizismus-Score erheblich gesunken.

Welche Interpretation ist angemessen?

- Tritt Fall 1 ein, so handelt es sich im Vor- und im Nachtest um „Neurotizismus“ ($r_{\text{prä, post}} \Rightarrow 1$). Diese Feststellung ist aber „diagnostisch wertlos“ - das Merkmal wurde völlig fehlerhaft erfaßt ($r_{\text{tt diff}} \Rightarrow 0$).
- Tritt Fall 2 ein, so wird Prä- und Posttest fehlerfrei erfaßt ($r_{\text{tt diff}} \rightarrow 1$). Nur handelt es sich beim Nachtest nicht mehr um dasselbe Merkmal „Neurotizismus“ wie beim Vortest ($r_{\text{prä, post}} \Rightarrow 0$).

Kasten 15-2 veranschaulicht das Dilemma an einem Beispiel.

Kasten 15-2:
Veranschaulichung des Reliabilitäts-Validitäts-Dilemmas

Quelle: Leichner, 1979, 50-57

Die Reliabilität von Differenzen (Messung 1 minus Messung 2) lässt sich durch folgende Gleichung schätzen:

$$r_{\text{tt diff}(A,B)} = \frac{r_{AA} + r_{BB} - 2 r_{AB}}{2 - 2 r_{AB}}$$

Es bedeuten:

- A : Messung vor der Intervention (Prätest),
- B : Messung nach der Intervention (Posttest),
- $r_{\text{tt diff}(A,B)}$: Reliabilität der Differenzen zwischen A und B,
- r_{AA} : Reliabilität des Prätests A,
- r_{BB} : Reliabilität des Posttests B,
- r_{AB} : Korrelation des Prätests A mit dem Posttest B.

Zwei Fälle sollen das Dilemma veranschaulichen!

Fall 1: Setzt man die Reliabilität von A und B gleich Eins, die Korrelation von A und B ebenfalls gleich Eins, dann folgt: Die Reliabilität der Differenzen fällt auf Null.

Sei: $r_{AA} = r_{BB} = 1$ und $r_{AB} = 1$.

Einsetzen:
$$r_{\text{tt diff}(A,B)} = \frac{1 + 1 - 2 \cdot 1}{2 - 2 \cdot 1} = 0$$

Fall 2: Setzt man die Reliabilität von A und B gleich Eins, die Korrelation von A und B dagegen gleich Null, dann folgt: Die Reliabilität der Differenzen steigt auf Eins.

Sei: $r_{AA} = r_{BB} = 1$ und $r_{AB} = 0$.

Einsetzen:
$$r_{\text{tt diff}(A,B)} = \frac{1 + 1 - 2 \cdot 0}{2 - 2 \cdot 0} = 1$$

Kommentar:

- In Fall 1 handelt es sich bei dem Ausgangszustand A und dem Endzustand B um dasselbe Merkmal ($r_{AB} = 1$). Aber ihre Erfassung ist diagnostisch wertlos, weil völlig fehlerhaft ($r_{\text{tt diff}(A,B)} = 0$).
- In Fall 2 ist die Messung völlig fehlerfrei ($r_{\text{tt diff}(A,B)} = 1$). Aber der Ausgangszustand A und der Endzustand B „haben nichts miteinander zu tun“ ($r_{AB} = 0$).

15.4 Zusammenfassung zu Kapitel 15

Der Diagnostiker kann Merkmale auf unterschiedlichen Ebenen untersuchen: auf der Ebene der Performanz mit Verfahren, die jene ‚Handlungen‘ evozieren, die ein Merkmal kennzeichnen; auf der Ebene der Deskription mit Verfahren, die ein Merkmal (nur) beschreiben. Nützlich erweisen sich Verfahren beider Ebenen vor allem bei komplexen Fragen.

Gegenstand diagnostischen Bemühens kann es sein, die gegenwärtige Situation eines Probanden zu erfassen oder aber seine biographische Entwicklung zu beschreiben.

Ebenso kann es darum gehen, nur einen Ist-Zustand zu ermitteln, man spricht von Statusdiagnostik, oder aber einen Verhaltens-Verlauf zu verfolgen, gegebenenfalls bis zu einem Soll-Zustand hin, man spricht von Prozeßdiagnostik.

Prozeßdiagnostik soll Veränderungen erfassen. Erfassung von Verhaltensänderungen wirft spezielle Fragen auf, die bisher noch nicht zureichend gelöst sind.

15.5 Kontrollfragen zu Kapitel 15

Diagnostik auf der Ebene der Performanz.

Diagnostik auf der Ebene der Deskription.

Probleme beider Vorgehensweisen.

Aktuelle Diagnostik.

Biographische Diagnostik.

Verfahren im Dienste einer Aktuellen Diagnostik.

Verfahren im Dienste einer Biographischen Diagnostik.

Statusdiagnostik.

Prozeßdiagnostik.

Konzept der Veränderungsmessung.

Modelle der Veränderungsmessung.

Probleme der Veränderungsmessung.

16. Kapitel

Erfolgskontrolle

Zur medizinischen Intervention gehört die Entscheidung darüber, ob der Klient ‚geheilt/zum Teil geheilt/nicht geheilt‘ sei. Gehört eine solche Evaluation auch zur psychologischen Diagnostik: eine Prüfung also, wie weit sich die ‚Antwort‘ auf die diagnostische Fragestellung bewährt (bewahrheitet) hat?

Diagnostik verursacht ‚Kosten‘ (Honorare/Arbeitszeit/Sachausgaben usw.): ‚Rentiert‘ sich dieser Aufwand? - Der Diagnostiker ruft psychologisches Wissen ab und wendet psychologische Techniken an: Haben sich ‚diagnostisches Urteil‘ und daraus abgeleitete ‚Interventionen‘ als ‚richtig‘ erwiesen im Sinne der Wissenschaft Psychologie? - Von (mindestens) zwei Seiten her rechtfertigt sich die Frage nach einer Evaluation: vom Forschungskontext und vom Sachaufwand her.

Ins Spiel kommen aber weitere Bewertungsaspekte (Göllner & Deter, 1980; Halder-Sinn, 1980, 92-93; Vennen, 1992; Wottawa & Hossiep, 1987, 90-100):

- Vorteil und Nachteil für den Probanden,
- Rückmeldungen an den Diagnostiker (Bestätigungs- oder Versagenserlebnisse),
- Nutzen und Schaden für gesellschaftliche Institutionen, für Bezugspersonen des Probanden.

Unter verschiedener Perspektive rechtfertigt sich demnach die **Frage nach dem Erfolg** diagnostischer (vor allem modifikatorischer) Bemühungen.

Zu ermitteln ist der ‚diagnostisch-interventive Effekt‘ - abgehoben von Effekten, die nicht auf Diagnostik oder Intervention zurückgehen, sondern auf andere (unbekannte?) Variablen.

Wie aber bemißt sich dieser Effekt? Durch Vergleich einer Ausgangs- mit einer Endsituation (gegebenenfalls auch eines Verlaufs)! Wieder, wie so oft, ist damit die Grundfrage nach Kriterien aufgeworfen.

Zu definieren sind

- Kriterien, die den *Ausgangszustand* erfassen,
- Kriterien, die den *Änderungsprozeß*, und
- Kriterien, die die *Endsituation* bestimmen.

Ein einheitlicher Kriterien-Katalog ist offensichtlich nicht erstellbar. In Analogie zu einer Liste bei Halder-Sinn (1980, 93) seien Beispiele genannt:

- Urteil des *Diagnostikers* oder *Therapeuten*,
- Selbstbeurteilung des *Probanden/des Klienten*,
- Urteil *unabhängiger Diagnostiker* oder *Therapeuten*,
- Urteil *unabhängiger Bezugspersonen* des Probanden (z. B. seiner Familienangehörigen, seiner Freunde),
- Daten aus *Leistungstests*,
- Daten aus *Persönlichkeitsinventaren*,
- Daten aus *projektiven Verfahren*,
- spezielle *interventionsorientierte Konzepte* (andere bei Schulpsychologie als bei Verhaltens- oder Gesprächstherapie).

Als Beispiel für Evaluation im Einzelfall sei das Vorgehen bei der Verhaltensdiagnostik/Verhaltenstherapie genannt. Ermittelt werden Ausgangsniveau (Basisrate), Therapieverlaufsplan und Endniveau (Endwert), so daß sich durch Vergleich immer angeben läßt, was (schon oder endgültig) erreicht worden ist (Lutz & Windheuser, 1976, 204).

In Studien erweist sich die Evaluation als sehr aufwendig. Entworfen werden experimentelle oder quasi-experimentelle Pläne, einbezogen werden Versuchs- und Kontrollgruppen. Das Beispiel in Kasten 16-1 kann dies veranschaulichen.

Kasten 16-1: Ein Beispiel für Erfolgskontrolle

*Wie aufwendig Erfolgskontrollen sind,
wenn sie inhaltlich und methodisch überzeugen sollen, belegt ein Beispiel
von Sloane et al. (1975a, 1975b), Staples et al. (1975, 1976).*

1. 94 neurotische Klienten wurden nach Zufall zwei **Therapiegruppen und einer Kontrollgruppe** (Warteliste) zugeordnet:
 - a) **Verhaltenstherapiegruppe:** Zugeordnet waren 30 Klienten, (auswertbar waren 28 Interview-Protokolle).
 - b) **Psychoanalysegruppe:** Zugeordnet waren 30 Klienten, (auswertbar waren 22 Interview-Protokolle).
 - c) **Kontroll-Gruppe:** Auf der Warteliste standen die restlichen 34 Klienten. In Gruppe a und b waren 6 Therapeuten tätig, 3 Verhaltenstherapeuten und 3 Psychoanalytiker. Jeder Therapeut hatte 10 Klienten. Die Therapie dauerte 4 Monate.
2. Zur Erfolgsmessung dienten **zwei Verfahrensklassen:**
 - a) Es wurden **standardisierte** Skalen verwandt. Zwei Beispiele:
 - die *Therapist Information-Specifity Scale* (Lennard & Bernstein, 1960: Diese Skala mißt die Interaktion von Therapeut und Klient);
 - das *Relationship Questionnaire* (Truax & Carkhoff, 1967: Diese Skala erfaßt die Einstellung des Klienten zum Therapeuten).
 - b) Es wurden **unstandardisierte** Verfahren eingesetzt. Zwei Beispiele:
 - *Sprechcharakteristiken des Klienten* wurden erfaßt, indem man Pausen oder Länge der Sprecheneinheiten maß.
 - Auf Ratings gaben die *Therapeuten ihre eigenen Gefühle* gegenüber den Klienten an und schätzten die Haltung des Klienten zu Therapie/ Therapeut ein.

3. Den Erfolg schätzten **unterschiedliche Urteiler** ein:
 - a) die *Klienten/Probanden* selber,
 - b) die behandelnden *Therapeuten*,
 - c) Kliniker, die an der Therapie *unbeteiligt* waren,
 - d) *Bezugspersonen* der Klienten (Familienangehörige).
4. Der Zustand der Klienten wurde **zu vier Meßzeitpunkten** eingeschätzt:
 - a) vor *Beginn* der Therapie,
 - b) zum *Abschluß* der Therapie,
 - c) *vier Monate nach* Abschluß der Therapie,
 - d) *zwölf Monate nach* Abschluß der Therapie.
5. Nach vier Monaten Therapie hatte sich bei 50 Prozent der Klienten aus der Kontroll- und bei 50 Prozent aus der Therapiegruppen eine ‚Besserung‘ eingestellt. Als Erfolg galt es, wenn sich Neurose-Symptome ‚gebessert‘ hatten, die zu Beginn der Therapie festgestellt worden waren.

Die Wirkung der beiden Therapieformen war insgesamt etwa gleich.

Im einzelnen gab es jedoch Unterschiede:

- *Verhaltenstherapeuten übertrafen die Psychoanalytiker* in drei Dimensionen: in
 - ⇒ Empathie,
 - ⇒ persönlichem Kontakt (interpersonal contact) und
 - ⇒ ‚Echtheit‘ (self-congruence).
- Nach dem Urteil der Klienten
 - ⇒ waren die *Verhaltenstherapeuten* autoritärer,
 - ⇒ ermutigten die *Psychotherapeuten* zu mehr *Unabhängigkeit*.
- Klienten, die *mehr sprachen*, wiesen einen *höheren Erfolgswert* auf als Klienten, die weniger redeten.

Zusammenfassung zu Kapitel 16

Zur Diagnostik als Wissenschaft gehört die Aufgabe, ihre Ergebnisse auch zu evaluieren. Im Einzelfall kann sich diese Aufgabe häufig als unrealisierbar erweisen.

Es gibt verschiedene Gründe, die eine Erfolgskontrolle rechtfertigen, beispielsweise die Kosten, die Diagnostik und Intervention verursachen, oder die Auswirkungen auf die Lebenssituation des Probanden.

Besondere Schwierigkeiten wirft die Auswahl valider Kriterien auf, an denen der Ausgangs- und der Endzustand gemessen werden sollen.

Kontrollfragen zu Kapitel 16

- Gründe für eine Erfolgskontrolle.
- Aspekte einer Erfolgsbewertung.
- Mögliche Kriterien einer Erfolgsmessung.

- Erfolgsmessung im Einzelfall.
- Erfolgsmessung im Dienste der Diagnostik als Wissenschaft.
- Schwierigkeit der Festlegung von Erfolgskriterien.

17. Kapitel

Nutzenschätzung: *Entscheidungstheorie und Diagnostik oder Intervention*

Ein diagnostisches Urteil ist immer entscheidungsorientiert. Entschieden oder wenigstens mit-entschieden wird beispielsweise durch Selektion über die Zukunft eines Bewerbers, durch Therapiezuweisung über die „Gesundheit“ eines Klienten.

Was das Konzept der ‚Entscheidungstheorie‘ jedoch in die Diagnostik einführen soll, zielt auf einen besonderen Aspekt der Entscheidung.

*Es geht **um** eine **Entscheidung**, die gefällt wird im Blick*

- auf den **Nutzen**, den eine diagnostische Arbeit stiftet, und
- auf die **Kosten**, welche sie verursacht.

Der entscheidungstheoretische Ansatz erweitert den Rahmen der klassischen Testtheorie. Unter welchem Aspekt?

- Die klassische Testtheorie bemißt den Wert von Verfahren nach dem Grade ihrer Objektivität, Reliabilität und Validität.
- Dagegen bewertet der entscheidungstheoretische Ansatz ein Verfahren unter dem Aspekt von Kosten und Nutzen: „Das Kriterium für den Wert eines Tests . . . ist nicht so sehr irgendein Grad der Genauigkeit, die er selber hat, vielmehr der Beitrag, den er für das Urteil leistet“ (Cronbach & Gleser, 1965, 148).

Die Frage darum lautet „nicht mehr: ‚Taugt der Test etwas?‘, sondern: ‚Taugt der Test etwas für meine Untersuchungszwecke?‘“ (Klapprott, 1975, 90).

*„Die Nützlichkeitsprüfung unterscheidet sich von der Validitätsbestimmung dadurch, daß sie spezifisch für jede Entscheidungssituation vorgenommen werden muß. Wesentlich dabei ist, daß man den Nutzen eines Tests in Beziehung zur besten a-priori-Strategie oder Grundrate zu setzen hat und ihn nur dann als gegeben ansehen kann, wenn mehr Information geliefert wird, als man aufgrund anderer Entscheidungsstrategien erhalten könnte; d. h. es empfiehlt sich, erst einmal zu sehen, welche Entscheidungen man **ohne** den Test treffen kann, um das Ergebnis mit der Entscheidung zu vergleichen, die mit Hilfe des Tests zu treffen ist“ (Schmidtchen, 1975, 34).*

Wir stellen Ansätze vor, die sich unterscheiden durch die Art der Nutzenschätzung (Wottawa & Hossiep, 1987, 42-51).

Nach den einzelnen Modellen wird der Nutzen unterschiedlich geschätzt, beispielsweise

- durch Zerlegung in Einzelkomponenten (17.1),
- durch Vergleich der Vorzüge verschiedener Methoden (17.2),
- durch Angaben in Geld (17.3),
- durch Befragung von Experten (17.4),
- durch Befragung der Betroffenen (17.5).

Es folgen eine Zusammenfassung (17.6) und die Vorgabe einiger Kontrollfragen (17.7).

17.1 Nutzenschätzung durch Zerlegung in Einzelkomponenten

Modell von Cronbach und Gleser

Cronbach und Gleser haben einen Algorithmus vorgeschlagen, der es erlauben soll, den Gesamtnutzen diagnostischer Untersuchungen zu schätzen, indem der Gesamtnutzen und die Gesamtkosten in einzelne Komponenten zerlegt werden (1965).

Die Formel lautet *vereinfacht*:

$$Nu_{ges} = Nu_e - Ko$$

Es bedeuten:

- Nu_{ges} : Gesamtnutzen,
- Nu_e : Nutzen von Einzelkomponenten,
- Ko : Kosten.

Demnach ergibt sich der **Gesamtnutzen** (Nu_{ges}) als Differenz zweier Größen: des Nutzens von *Einzelkomponenten* (Nu_e) und des *Kostenaufwandes* (Ko).

Der **Nutzen von Einzelkomponenten** (Nu_e) ergibt sich als ein *Produkt*, in das eingehen

- der Informationsnutzen eines *Scores* (aus Test, Exploration oder projektivem Verfahren usw);
- der Behandlungsnutzen, den ein *Treatment* liefert;
- der Ergebnisnutzen, der an einem *Kriterium* zu messen ist.

Die **Kosten** ergeben sich als Summe aller Elemente, die zu den Kosten beitragen.

Berechnet werden Einzelnutzen und Kosten für die Gesamtzahl der Probanden (N). Wenn N mit aufgenommen wird, lautet die Formel zur Schätzung des Gesamtnutzens - *wieder vereinfacht*:

$$Nu_{ges} = N (\sum Nu_X \cdot \sum Nu_T \cdot \sum Nu_C) - N \sum Ko$$

Es bedeuten:

Nu_{ges} : Gesamtnutzen,
 Nu_X : Informationsnutzen eines einzelnen Scores X,
 N : Behandlungsnutzen eines Treatments T,
 Nu_C : Ergebnisnutzen, gemessen an einem Kriterium C,
 Ko : Kosten.

In beide Größen (Einzelnutzen und Kosten) gehen die **Auftretenswahrscheinlichkeiten** der einzelnen Determinanten als Gewichte mit ein (hier in der Formel zur Vereinfachung weggelassen).

Den Nutzen einer diagnostischen Untersuchung zu schätzen, wie Cronbach & Gleser vorschlagen, setzt voraus, daß sich jede der Größen genau bestimmen läßt und **alle beteiligten Größen in vergleichbaren Einheiten** angegeben werden. Eben darin liegt jedoch die Schwierigkeit: Wie soll beispielsweise der Nutzen geschätzt werden, der einer einzelnen Information zukommt? Wie der Nutzen, der einer Behandlung entspringt?

Die Beispiele, in denen das Modell von Cronbach & Gleser vorgestellt wird, beruhen auf **angenommenen** Nutzen- und Kostengrößen (Gösslbauer, 1981; Michel & Mai, 1968).

17.2 Nutzungsschätzung durch Vergleich der Vorzüge verschiedener Methoden „Multiattributive Nutzenschätzung“

Im Modell der sogenannten „Multiattributiven Nutzenschätzung“ wird die Gesamtentscheidung aufgegliedert in Teilentscheide - ähnlich wie bei Cronbach und Gleser.

*Doch wird **explizit nur der Nutzen** geschätzt. Verglichen werden die Vorzüge verschiedener Methoden. Vorrang erhält jene Methode, deren Nutzen am höchsten bewertet wird. -Implizit geht in den Vergleich auch die Abschätzung der Kosten ein.*

Für jede Methode wird der Gesamtnutzen in Segmente zerlegt, jedes Segment dann isoliert bewertet. Ein Algorithmus fügt die Einzelbewertungen additiv zu einer Gesamtschätzung zusammen.

Eine formalisierte Methode dieser **Art** ist die **multiattributive Nutzenschätzung** von Edwards, W. (1980: Multi-Attributive-Utility-Theory: MAUT).

„Das Prinzip der MAUT ist die Zergliederung des Entscheidungsproblems in einzelne Teile, die jedes für sich relativ leicht zu beurteilen sind, und

die anschließende Zusammenfügung dieser Teilurteile in einem formalen Modell“ (Kasubek & Aschenbrenner, 1978, 595).

Kasten 17-1 soll veranschaulichen, wie sich die Gesamtentscheidung in Teilschritte aufgliedern läßt.

Kasten 17-1:
Multiattributive Nutzentheorie:
Aufgliederung einer Entscheidung in Teilentseide

„Im additiven Modell ergibt sich der Gesamtwert einer Alternative als gewogene Summe der Teilnutzwerte ihrer Attributsausprägungen“ (Kasubek & Aschenbrenner, 1978, 598). Es gilt die Gleichung:

$$U(A_j) = W_1 U_1(X_{1j}) + W_2 U_2(X_{2j}) + \dots + W_n U_n(X_{nj})$$

Es bedeuten:

$U(A_j)$: Gesamtnutzen der Alternative j,

X_{1j} : Ausprägung der Alternative j auf dem Attribut 1,

U_1 : Teilnutzwert von X_1 für Alternative 1,

W_1 : Gewicht des Attributes 1.

Beispiel: Vorgegeben seien

- als *Alternativen* sechs Therapieformen,
- als *Attribute* die Effektivität der verschiedenen Therapieformen bei sechs Arten von Verhaltensstörungen,
- als *Ausprägung* ein Rating von 1 bis 6 (geringe bis große Effektivität),
- als *Teilnutzwert* der Beitrag der Ausprägung X zur Effektivität auf dem jeweiligen Attribut, angebar auf einem Rating von 1 bis 6 (geringer bis großer Nutzen),
- als *Gewicht* die Bedeutsamkeit einer Alternative (Therapieform) je Attribut (Effektivität), angebar wiederum auf einem Rating von 1 bis 6 (geringe bis große Bedeutsamkeit).

Fünfzehn Therapeuten mit mehr als fünf Jahren Berufserfahrung werden gebeten, die sechs Alternativen zu bewerten.

Ihre Bewertungen gehen in die Gleichung ein und erlauben es, den „Gesamtnutzen“ je Alternative zu schätzen und auf diese Weise die sechs verschiedenen Therapieformen zu vergleichen.

Vorteil: Die Nutzen- und Kostenaspekte werden einzeln gewichtet, unabhängig von der Gesamtichtung, die Einzelwichtungen dann durch einen Algorithmus zusammengefaßt.

„Eine wesentliche Stärke des Verfahrens ist es zweifellos, daß inhaltskompetente Personen angeleitet werden, Entscheidungsprobleme zu strukturieren, bis hin zu operationalisierbaren Konsequenzen zu durchdenken und auf zentrale Merkmale und Merkmalsdimensionen zu verdichten“ (Miüller, G. F. & Nachreiner, 1988, 126).

Schwierigkeit: Läßt die Praxis den Vergleich von alternativen Vorgehensweisen zu? Verfügt ein Praktiker immer über alternative Vorgehensweisen, deren Nutzen er vergleichen kann?

„Trotz der Vorteile dieses Verfahrens kann nicht verkannt werden, daß es auf Prämissen aufbaut, die faktisch im klinischen Bereich nicht gegeben

sind: Die prinzipielle Verfügbarkeit von Behandlungsarten auf seiten des Psychotherapeuten.“ Psychotherapeuten wenden „jene Verfahrensweisen“ an, „die sie gelernt haben bzw. in denen sie durch ihr therapeutisches Agieren bestätigt werden“ (Jäger; R. S., 1986, 288).

„Eine Schwäche der multi-attributiven Nutzentechnik ist es, daß sie mit Expertenwissen arbeitet, dessen Vorhersagegültigkeit selbst jedoch nicht kontrolliert wird... Eine weitere Schwäche des Verfahrens birgt die u. U. inadäquate Messung von Merkmalsausprägungen. Probleme können sowohl bei der Operationalisierung als auch bei der Nutzenbeurteilung auftreten“ (Müller G. F. & Nachreiner 1988, 126-127).

17.3 Nutzenschätzung durch Angaben in Geld

Die Kosten einer diagnostischen Untersuchung und einer interventiven Maßnahme werden in Geld angegeben. Dieser Schritt ist eine einfache Prozedur, wenn realitätsnahe Schätzungen möglich sind.

„Am einfachsten ist es, wenn sich die relevanten Kriterien einfach in Geld umrechnen lassen, etwa bei den Kosten für eine abgebrochene Ausbildung oder den Aufwendungen für die neue Besetzung einer zunächst mit einer ungeeigneten Kraft besetzten Stelle. Diesem Vorgehen sind aber meistens enge Grenzen gezogen, da sich subjektive Komponenten (z.B. die Enttäuschung durch den Ausbildungsabbruch, der psychische Streß von Mitarbeitern als Folge des Fehlverhaltens ungeeigneter Vorgesetzter) damit oft nicht adäquat erfassen lassen“ (Wottawa & Hossiep, 1987, 42).

Vorteil: Nutzen und Kosten werden in derselben ‚Einheit‘ geschätzt.

Schwierigkeit: Lassen sich alle Nutzen- und Kostenfaktoren in Geld angeben?

17.4 Nutzenschätzung durch Experten

Wenn Experten den Nutzen schätzen sollen, bieten sich mehrere Möglichkeiten an. Drei Varianten seien genannt:

- Gruppendiskussionen zwischen Experten,
- Delphi-Methoden,
- Szenario-Techniken,

Gruppendiskussionen zwischen Experten

Bei der Gruppendiskussion wird ein diagnostisch-interventives Problem vorgegeben, über das Experten diskutieren, um Nutzen und Kosten zu schätzen.

Vorteil und Schwierigkeit: Bekanntlich entfalten Gruppendiskussionen eine eigene Dynamik. Zwar kann diese Gruppendynamik die beteiligten Personen ‚aus ihrer Reserve‘ locken und sie so zu einer umfassenden Diskussion bewegen, aber sie kann auch dazu führen, daß sich nicht das kompetenteste Urteil durchsetzt, sondern die Meinung des sozial stärksten Experten.

„Es kann in dieser Situation schwer sein, bei Abwägung des Nutzens einer etwas erhöhten Treffsicherheit gegenüber einer Vervielfachung der Kosten der Diagnoseerstellung einen sachgerechten Kompromiß zu finden“ (Wotawa & Hossiep, 1987, 43).

Delphi-Methode

Die Schwierigkeiten, die einer Gruppendiskussion entspringen können, soll die sogenannte Delphi-Methode vermeiden.

Definition: *„Delphi may be characterized as a method for structuring a group communication process so that the process is effective in allowing a group of individuals, as a whole, to deal with a complex problem“ (Linstone & Turoff, 1975, 3).*

An der Delphi-Methode beteiligen sich zwei Gruppen:

- erstens eine Gruppe von Experten, welche das anstehende Problem bewertet: das *Experten-Team*, und
- zweitens eine Gruppe, welche den Informationsaustausch zwischen den Experten steuert: das *Monitor-Team*.

Experten-Team: Die Experten sitzen einander nicht von ‚Angesicht zu Angesicht‘ gegenüber, sondern jeder sitzt - bildlich gesprochen - allein in einer ‚Zelle‘ und gibt seine Urteile ab, ohne die Urteile der Kollegen zu hören. (Die ‚Zelle‘ von Experte A kann in Hamburg, die von Experte B in Straßburg, die von Experte C in Baltimore liegen.) Jeder Experte wird schriftlich nach seiner Meinung befragt und leitet seine Urteile schriftlich dem Monitor-Team zu.

Monitor-Team: Das Monitor-Team faßt die Antworten der Experten zusammen und gibt die Zusammenfassung an die Experten zurück, die auf diesem Wege die Möglichkeit erhalten, ihre Ersturteile zu revidieren.

Regelkreis: Je nach Problemtyp, nach verfügbarer Zeit und nach den finanziellen Mitteln kann dieser ‚Regelkreis‘ des Informationsaustausches drei-, vier- oder fünfmal durchgespielt werden.

Computer-Einsatz: Das Monitor-Team kann sich unterstützen lassen durch Computer und durch Computernetze. Computer können die Experten-Antworten zusammenstellen, über Netz an die Experten zurückmelden und so den Informationsaustausch beschleunigen.

Vorteile: Das Urteil vieler Experten läßt sich leichter zusammenfassen als in der Gruppensituation. Vermieden werden die Informationsverzerrungen, die einer Gruppendiskussion entspringen können (Wottawa & Hossiep, 1987, 44).

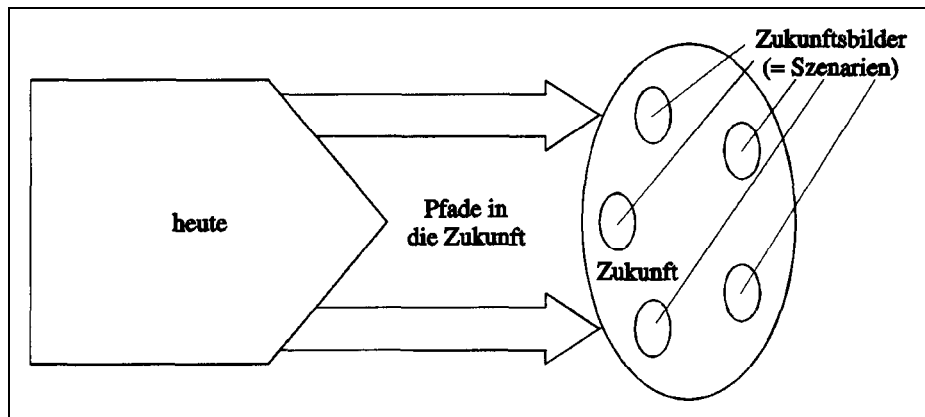
Schwierigkeit: „Das Verfahren ist aufwendiger und nimmt mehr Zeit in Anspruch als die Gruppendiskussion“ (Wottawa & Hossiep, 1987, 44).

Szenario-Technik

Die Szenario-Technik besteht darin, daß Zukunftsperspektiven (Szenarien) entworfen und diskutiert werden (Reibnitz, 1983).

Definition: „Unter einem Szenario versteht man sowohl

- die Beschreibung einer möglichen zukünftigen Situation als auch
- das Aufzeigen des Entwicklungsverlaufs, der zu dieser zukünftigen Situation hinführt.



Die Szenario-Technik ist demnach eine systematische Methodik zum Entwickeln von Szenarien; sie besteht aus acht logisch aufeinander aufbauenden Schritten, die den gesamten Prozeß transparent und in allen Phasen nachvollziehbar machen“ (Reibnitz, 1983, 112).

Die acht Schritte gibt Kasten 17-2 wieder.

Kasten 17-2:
Acht Schritte der Szenario-Technik
Quelle: Reibnitz (1983, 115-116)

1. Schritt: *Strukturierung und Definition des Untersuchungsfeldes,*
2. Schritt: *Identifizierung und Strukturierung der wichtigsten Einflußbereiche auf das Untersuchungsfeld (Umfelder),*
3. Schritt: *Ermittlung von Entwicklungstendenzen und kritischer Deskriptoren für die Umfelder*
4. Schritt: *Bildung und Auswahl konsistenter Annahmenbündel,*
5. Schritt: *Interpretation der ausgewählten Umfeldszenarien,*
6. Schritt: *Einführung und Auswirkungsanalyse signifikanter Störereignisse,*
7. Schritt: *Ausarbeiten der Szenarien bzw. Ableiten von Konsequenzen für das Untersuchungsfeld,*
8. Schritt: *Konzipieren von Maßnahmen und Planungen.*

Vorteil und Schwierigkeit: In die Entwicklung eines Szenarios gehen unvermeidbar viele subjektive Bewertungen mit ein. Dieses Wissen erzeugt bei den beteiligten Personen ein Unsicherheitsgefühl, kann aber gleichzeitig auch große Vorsicht wecken bei Beurteilung diagnostisch-interventiver Maßnahmen (Wottawa & Hossiep, 1987, 44)

17.5 Nutzenschätzung durch die Betroffenen

Der Nutzen diagnostischer und interventiver Maßnahmen läßt sich auch bestimmen durch die Betroffenen selber.

Dies besagt: Die betroffenen Probanden gewichten die vorgesehenen Schritte, beispielsweise den Wert eines Berufsweges, zu dem ein Psychologe rät. Doch setzt diese Gewichtung voraus, daß die Probanden angemessen informiert sind - beispielsweise durch ausführliche Diskussion mit ‚Experten‘. Sonst besteht die Gefahr, daß die Urteile der betroffenen Probanden zuvielen subjektiven Informationsverzerrungen unterliegen.

Welche Möglichkeiten bieten sich für solche Diskussionen an?

- Der Proband wird gründlich **informiert** und trifft **als Einzelner** seine Entscheidung.
- **Proband und Psychologe konzipieren gemeinsam** die Ziele bestimmter diagnostischer und interventiver Prozeduren. Die Dimensionen der **Zielerreichung** können sie sogar in einer Art Skala formulieren, an der sie den Erfolg der vorgesehenen Schritte messen (Kirusek & Sherman, 1968: goal-attainment-scale).
- Probanden und Psychologe bilden gemeinsam sogenannte **Planungszellen**: Kleingruppen, die sich über geplante Maßnahmen umfassend informieren, ihren Effekt abschätzen, ihren Wert diskutieren, schließlich darüber abstimmen, welche Maßnahmen verwirklicht werden sollen (Dienel, 1978).

Vorteil und Schwierigkeit: Die Betroffenen erhalten die Möglichkeit, ihr Wissen und ihre Vorstellungen in die Urteile über diagnostische und interventive Schritte einzubringen. Doch setzt dieser Prozeß voraus, daß Psychologe und Proband bereit sind, ein erhebliches Maß an Zeit aufzuwenden für den Informationsaustausch. Dabei müssen beide Partner (Psychologe und Proband) immer wieder auch die Frage stellen, wie valide die „subjektiven Einstufungen“ am Ende sind.

Resümee zum Konzept der Nutzenschätzung

Was leisten entscheidungstheoretische Modelle für die Lösung praktischer Probleme in Diagnostik und Intervention?

Generell gilt, „daß die Anwendung der Entscheidungstheorie in der Praxis dazu zwingt, alle Schritte des Entscheidungsprozesses explizit zu planen, festzulegen und darüber Rechenschaft abzulegen“ (Gösslbauer, 1981, 242).

Speziell gilt: „Der Gesichtspunkt der Nützlichkeit ist vor allem dann bedeutsam, wenn größere Testuntersuchungen geplant und durchgeführt werden sollen, also wenn eine Gruppe von Individuen mit Hilfe einer ganzen Batterie von Tests untersucht werden soll“ (Klapprott, 1975, 71).

Offenkundig ist auch: Zwar kann ein Diagnostiker oder Therapeut nur wenige der skizzierten Modelle anwenden - *solange er allein auf sich gestellt ist*. Dennoch kann die Vertrautheit mit entscheidungstheoretischen Ansätzen seinen Sinn dafür schärfen, diagnostisch-interventive Entscheidungen *auch* zu orientieren am sozialen oder individuellen Nutzen und an den anfallenden Kosten - *nicht allein* an den klassischen Gütekriterien.

17.6 Zusammenfassung zu Kapitel 17

Die diagnostische Arbeit sollte sich nicht allein an den psychometrischen Gütekriterien diagnostisch-interventiver Verfahren orientieren, sondern auch an einer Schätzung des Nutzens, den ein Verfahren für Proband und Mitwelt bringt.

Der Nutzen läßt sich auf unterschiedliche Weise schätzen, beispielsweise

- durch Angaben in Geld,
- durch Befragung von Experten,
- durch statistische Modelle,
- durch Befragung betroffener wohl-informierter Probanden.

17.7 Kontrollfragen zu Kapitel 17

- Entscheidungstheorie und klassische Testtheorie.
- Nutzenschätzung nach Cronbach und Gleser.
- Nutzenschätzung in Geld.
- Nutzenschätzung durch Experten.
- Nutzenschätzung durch statistische Modelle wie die MAUT.
- Nutzenschätzung durch die Betroffenen.
- Vorteil und Schwierigkeit des entscheidungstheoretischen Ansatzes.

18. Kapitel

Computerdiagnostik

Eftychia Sidiropoulou

Die Diagnostik erfordert viele instrumentelle Handlungen. Einen Teil davon können Apparate erleichtern oder dem Diagnostiker sogar abnehmen - diese Rolle fällt heute vorrangig dem Computer zu.

Seit mehr als drei Jahrzehnten setzen Psychologen Computer ein. Als die Technologie des Mikrocomputers weiterentwickelt wurde, verbreitete sich seine Nutzung in nahezu allen anwendungsorientierten Disziplinen der Psychologie.

Eine Fülle von Programmen wurde entwickelt, welche einzelne Phasen der Testvorgabe dem Computer überließen, etwa die Instruktion und die Auswertung, in manchen Fällen sogar Interpretation, Befunderstellung und Stellungnahme. Diese Entwicklung beschreibt Kisser (1986) als Übergang von einem computer-unterstützten zu einem Computer-gesteuerten Testen.

Kapitel 18 behandelt folgende Thematiken:

- Einsatz des Computers in der psychologischen Test-Diagnostik (18.1),
- Computersysteme (18.2),
- Einsatzfelder für eine computergestützte Diagnostik (18.3).

Im Anschluß an diese Teilkapitel wird die Problematik der *Äquivalenz zwischen Papier-Bleistift-Tests und ihrer Computer-Versionen* im Rahmen eines **Exkurses** dargelegt.

Das Kapitel schließt mit einer Zusammenfassung (18.4) und einer Reihe von Kontrollaufgaben (18.5).

18.1 Einsatz des Computers in der psychologischen Test-Diagnostik

Klieme und Stumpf (1990) haben in einer tabellarischen Übersicht aufgelistet, in welchen Phasen der Konstruktion und Anwendung von Tests der Computer

eingesetzt wird. Jäger, R. S. (1990) sowie Jäger, R. S. und Krieger (1994) haben diese Übersicht erweitert. Kasten 18-1 gibt die erweiterte Version wieder.

Kasten 18-1:
**Einsatz des Computers in einzelnen Phasen der Konstruktion
 und Anwendung von Tests**

Quelle: Jäger, R.S. und Krieger (1994, 218)

Testentwicklung:

- Itemgenerierung und Testkonstruktion
- Anlage und Verwaltung von Item-Banken
- Erstellung des Testmaterials
- Testerprobung und erste Revision

Testevaluation:

- Item- und Testanalysen
- Auswertung von Validierungsdaten
- Revision von Test- oder Testbatterien
- Anlage von „Normbanken“

Testdurchführung:

- Auswahl des zu bearbeitenden Tests
- Itemauswahl
- Testinstruktion/Testübungsphase
- Itempräsentation, eingeschlossen Simulation und interaktives Video
- Registrierung der Antwort und zusätzlich anfallender Daten

Testauswertung:

- Scoring
- normbezogene/kriterienbezogene Auswertung
- Analyse von Profilen oder Testverläufen
- Einzelfallstatistiken
- Rückmeldung an den Diagnostiker und an den Probanden (graphisch, numerisch, verbal)
- Interpretation von Testergebnissen
- Urteilsbildung und Indikation
- Angebote von Entscheidungshilfen und Expertensystemen
- Dokumentation, Speicherung von Daten
- Einzelfallkontrolle im Sinne der „Kontrollierten Praxis“
- Gutachten-Erstellung
- Gutachten-Validierung

Testentwicklung

Der Computer bietet die Chance, einen Test ökonomisch zu entwickeln; denn er ermöglicht es, Items automatisch zu generieren.

Voraussetzung dafür ist allerdings, daß dem Computer präzise Regeln ‚vorgegeben‘ werden, an denen sich die Itemkonstruktion orientiert. Gelingt es, solche präzisen Regeln zu ‚formulieren‘, kann der Computer beispielsweise für einen einfachen Rechentest neue Items generieren.

Diese Methode kann als Versuch angesehen werden, die *Inhaltsvalidität* eines Tests schon in der Konstruktionsphase zu erhöhen - die Konstruktionsalgo-

rithmen werden vom Inhalt und vom Aufbau der Items her entworfen. Insbesondere die Pädagogische Psychologie hat diese Möglichkeit genutzt. Es geht bei dieser Validierung nicht um eine empirische Erprobung, etwa durch Vergleich der Items mit einem Kriterium, sondern um inhaltliche Gruppierung der Items (Hornke & Rettig, 1989; Klieme & Stumpf, 1990).

Inhaltlich zusammengehörige Items können in sogenannten *Item-Banken* gesammelt werden, wenn sie eine homogene Skala abgeben: Das Merkmal der Homogenität ist dabei immer zu verstehen im Sinne eines bestimmten Modells, etwa der klassischen oder der probabilistischen Testtheorie.⁹

In den Item-Banken wird der Text eines Items gespeichert, ebenso wichtige Parameter, etwa Schwierigkeits- oder Diskriminationsindizes (Jäger, 1990; Klieme & Stumpf, 1990). Wenn Items sich als homogen erweisen und ihre Parameter geschätzt sind, können sie zur *Erstellung des Testmaterials* verwandt werden. Beispielsweise bieten sie sich an zur Erstellung von Paralleltests oder zur Konstruktion antwortabhängiger, sogenannter adaptiver Testverfahren. (Dazu siehe Kapitel 18.2.1.3, S. 391!)

Die Zusammenstellung der Items erfordert spezielle Software-Programme, die eine graphische Ausgestaltung und bildliche Darstellung am Bildschirm ermöglichen. Dafür liegen neben den „Textverarbeitungs- und Desk-Top Publishing“-Programmen auch interaktive Software-Programme vor, sogenannte Autorensysteme, z. B. EIDOS (Warzecha, 1989), QUOTEX (Kuliga, 1990). Einen Überblick bieten Klieme und Stumpf (1990).

Testevaluation

Für die Testevaluation stehen verschiedene Computerprogramme zur Verfügung, wie SPSS (Statistical Package for the Social Science: siehe Norusis, 1986) SAS (Statistical Analysis System: Institute Inc. Cary, 1983) oder CSS (Complete Statistical System: StatSoft. Inc., 1986-1991), welche die Item- und Testanalysen vornehmen, die Reliabilitäts- und Validitätskennwerte errechnen oder sogar -über bestimmte Unterprogramme- „Normbanken“ anlegen können.

Testdurchführung

Für die Testdurchführung erschließt der Computer neue diagnostische Möglichkeiten. Insbesondere können mit dem Computer adaptive Teststrategien realisiert werden: Adaptive oder antwortabhängige Teststrategien umschreiben

⁹ In den Testtheorien hat sich keine einheitliche Auffassung von Homogenität durchgesetzt, bei allen Unterschieden gilt aber, daß sie den Grad angibt, in dem die Items eines Tests dieselbe Eigenschaft messen. Zwei vereinfachte Beispiele: In der klassischen Testtheorie gelten Items als homogen, wenn sie hoch miteinander korrelieren. In den probabilistischen Testtheorien gelten Items als homogen, wenn ihre Verlaufskurven gleichartig sind.

ein Vorgehen, das der Proband mit seinem Antwortverhalten mitsteuert. „Adaptiv“ kann auch die *Instruktions- oder die Testübungsphase* dargeboten werden (siehe Abschnitt 18.2.1.3, S. 391).

Testinstruktion: Erwähnt sei das Konzept der „maßgeschneiderten Instruktionen“, in dem sich zwei Anliegen verbinden: Erstens soll dem Probanden das Verständnis der Aufgabenstellung vermittelt, zweitens aber die Informationszeit minimalisiert werden.

Dabei geht es um „lineare Lernprogramme, die ein schrittweises Erlernen der Aufgabenstellung in aufsteigender Komplexität beinhalten“ (Bukasa, Kisser & Wenninger, 1990, 151). Solche Programme ermöglichen es dem Probanden, selber die Bearbeitungszeit zu bestimmen. Bukasa, Kisser und Wenninger (1990) berichten, daß solche Instruktionsprogramme die Gesamttestdauer erheblich verkürzt hätten im Vergleich zu herkömmlichen Gruppentests.

In diesem Zusammenhang wendet Kubinger ein, Papier-Bleistift-Tests böten die Instruktionen sowohl akustisch als auch visuell (Testleiter liest vor/Proband liest mit), ein Computertest dagegen präsentiere die Instruktion nur visuell (am Bildschirm), so daß „auditive Wahrnehmungstypen benachteiligt“ seien (1993, 134).

Nur im Rahmen des Psychologischen Dienstes der Deutschen Bundeswehr (Wildgrube, 1990), sowie des Bundesanstaltes für Arbeit innerhalb des Dezentralen Testvorgabe- und Auswertungssystems (DELTA: siehe Hilke, 1993) wird über eine auditive Unterstützung (durch Kopfhörer) der Instruktionen berichtet.

Testübungsphase: Ähnlich wie die Instruktion kann auch die Testübungsphase auf den individuellen Arbeitsstil zugeschnitten werden. Dem Probanden bleibt es überlassen, ob er alle oder nur einen Teil der Übungs-Items bearbeitet.

Eine adaptive Darbietung der *Testinstruktion und Testübungsphase* könnte die klassische Testtheorie nicht zulassen. Mehrere theoretische Voraussetzungen wurden verletzt: Verletzt wurde erstens die Testobjektivität, nach der die Testbedingungen für alle Probanden gleich (standardisiert) sein müssen. Verletzt wurde zweitens die Reliabilität, nach der die Variation von Testbedingungen die Replizierbarkeit der Testergebnisse vermindert; dies gilt insbesondere für den Fall der Retest-Reliabilität.

Probabilistische Tests erlauben dagegen eine adaptive Testdarbietung (siehe Kap. 18.2.1.3, S.391).

Itempräsentation: Der Computer kann die Itempräsentation verbessern. Durch die Graphikfähigkeit des Computers ist es möglich, Objekte beweglich darzustellen und mit akustischen Signalen zu kombinieren (siehe Abschnitt 18.2.2, S. 395).

Der Computer ermöglicht *es, zusätzliche Daten* zu registrieren, etwa die Korrektur von Antworten, die Anzahl gelöster Items, die Latenzzeiten je Item. (Als Latenzzeit sei verstanden das „Intervall vom Beginn der Darbietung eines Items auf dem Bildschirm bis zum Moment der Registrierung einer Reaktion durch das Gerät“ [Klieme & Stumpf, 1990, 36].) Zur Zeit werden zusätzliche Daten allerdings noch nicht in die diagnostische Auswertung aufgenommen (Jäger, R.S., 1990; Wildgrube, 1990).

Auswertung

Der Beitrag des Computers bei der Verwertung und Bearbeitung von Test-Scores ist schon eine Selbstverständlichkeit.

- Er zählt und ordnet die Rohwerte.
- Er errechnet Standardwerte, erstellt Profile und vergleicht sie.
- Er schätzt Personen- und Itemparameter.
- Er registriert zusätzliche Daten, etwa Antwort-Korrekturen oder Latenz-Zeiten.
- Er stellt Profile numerisch, verbal oder graphisch dar.

Darüber hinaus besteht die Möglichkeit einer „automatisierten Rückmeldung“ an den Probanden. Diesbezüglich sei jedoch folgendes angemerkt: sie ist in der Regel kurz gefaßt, sie kann keine Rückfragen des Probanden klären. Darum bleibt das Gespräch mit dem Diagnostiker unverzichtbar.

Interpretation: Der Computer hat sich neuerdings auch bei Aufgaben durchgesetzt, von denen bisher galt, daß sie allein dem Psychologen vorbehalten seien. Genannt seien Aufgaben wie die folgenden:

- die Interpretation von Ergebnissen,
- die Erstellung von Gutachten,
- die Stellungnahme und die Indikationsangabe.

Die Komplexität solcher Aufgaben erschwert allerdings eine Umsetzung in Computerprogramme. Erste Versuche bestehen in der Entwicklung sogenannter Interpretationssysteme, vor allem aber sogenannter Expertensysteme. Ihre „Wissensbasis“ scheint eine individualisierte Datenverarbeitung zu „versprechen“ (siehe Abschnitt 18.2.4, S. 399).

Ein solcher Versuch kann nur im Kontext einer normativen Diagnostik gelingen. Zur Zeit wäre ein solcher Versuch unrealisierbar oder nur zu verwirklichen im Kontext „gemäßigt normativer Modellvorstellungen“. Ausführlich stellt Hageböck die Problematik dar (1994).

Schließlich ist festzuhalten: Der Computer soll die diagnostische Arbeit des Psychologen unterstützen, darf sie nicht ersetzen.

18.2 Computersysteme

Die praktische Umsetzung der Computerdiagnostik erfolgt durch die Planung und Entwicklung von *Computersystemen*. Unter diesem Begriff werden alle systematischen „Versuche“ zusammengefaßt, den Computer zu nutzen, um

- Papier-Bleistift-Tests vorzugeben,
- originale Computertests zu entwickeln und
- darüber hinaus komplexe diagnostische Systeme zu konzipieren.

Zur Zeit wird die Computerdiagnostik zum größten Teil durch die Übertragung und Durchführung von Papier-Bleistift-Tests auf den Computer betrieben. Dabei werden die vielfältigen technischen Möglichkeiten, die der Computer bieten kann, nicht in vollem Umfang genutzt.

Eine adäquate Nutzung des Leistungspotentials von Computern macht allerdings die Entwicklung einer Theorie computergestützter Diagnostik notwendig.

Dies kann an einem Beispiel verdeutlicht werden: Wie im vorangegangenen Abschnitt erwähnt, ermöglicht der Computer es, zusätzliche Daten (etwa Latenzzeiten, Bearbeitungszeiten usw.) simultan zu erheben. Die Bedeutung solcher Daten bleibt aber momentan nebensächlich. „Es fehlt“ nämlich „eine psychologische Theorie, die es erlaubt, solche Daten begründet abzuleiten und zu analysieren“ (Jäger, R. S. & Krieger, 1994, 220).

Vor dem Hintergrund der Entwicklung einer psychologischen Theorie haben Jäger, R. S. und Krieger (1994) versucht, ihre grundlegenden Elemente zusammenzustellen. Als Elemente gelten beispielsweise eine *inhaltliche Theorie* (z. B. Intelligenztheorie), welche um die Berücksichtigung der oben erwähnten zusätzlichen Daten erweitert werden soll oder eine *Theorie der Schnittstelle*, die sich mit denjenigen Faktoren befaßt, welche die Wechselwirkung zwischen Mensch und Computer betreffen (z.B. Maus).

Eine Nutzung des Computers in vollem Umfang bleibt allerdings zukünftiger Entwicklung vorbehalten. Dies macht den gegenwärtigen Tatbestand nicht weniger interessant: Diagnostik mit Computer stellt zur Zeit ein Forschungsfeld dar, auf dem auch neue Ansätze erprobt werden.

Im folgenden wird versucht, einen möglichst umfassenden Überblick über das Instrumentarium der Computerdiagnostik, nämlich die verschiedenen Computersysteme, zu geben.

Das Teilkapitel 18.2 gliedert sich in vier Abschnitte:

- Computertests (18.2.1),
- Computer-Testsysteme, Computer-Testgeräte (18.2.2),
- Computer-Interpretationssysteme (18.2.3),
- Computer-Expertensysteme (18.2.4).

18.2.1 Computertests

Es werden drei Varianten von Computertests dargestellt:

- Computer-Versionen von Papier-Bleistift-Tests (18.2.1.1),
- Computer-Simulationstests und (18.2.1.2),
- Adaptive Tests (18.2.1.3).

18.2.1.1 Computer-Versionen von Papier-Bleistift-Tests

Computerdiagnostik greift zur Zeit vorrangig zurück auf herkömmliche Papier-Bleistift-Tests. Aus Papier-Bleistift-Tests werden Computer-angepaßte Testversionen erstellt.

Die Übertragung von Papier-Bleistift-Tests zielt auf folgendes ab: Zum einen soll die Testbearbeitung in höherem Grade standardisiert, zum anderen die Zeit der Testdurchführung verkürzt werden.

Andererseits ist die Durchführung von Papier-Bleistift-Tests am Computer kritisch zu bewerten. Ausgehend von der Annahme, daß beide Versionen äquivalent sind, werden

- die Reliabilitäts- und Validitätsindizes, erhoben für die Papier-Bleistift-Version, ungeprüft auf die Computer-Version übertragen und
- die Normen der Papier-Bleistift-Version unverändert auf die Computer-Version angewandt.

Da die beiden Versionen mit verschiedenen Medien arbeiten (Papier oder Computer), ergeben sich veränderte Rahmenbedingungen der Testung. Die Gleichwertigkeit beider Versionen ist somit zu prüfen.

Am Beispiel der Itempräsentation läßt sich die Änderung der Rahmenbedingungen ausdrücklich aufzeigen. Die vielfältigen technischen Möglichkeiten des Computers (etwa Farbmonitore, CD-ROM) erlauben eine flexible Präsentation der Testaufgaben. Grob unterscheidet sie sich von der Papier-Bleistift-Vorgabe in folgenden Aspekten:

- *Format:* z.B. Anwendung von bewegten Bildern.
- *Antworteingabe:* Der Proband kann seine Antwort eingeben mithilfe eines Lichtgriffels oder spezieller Tastaturen.
- *Generelle Handhabung:* Darunter sind Möglichkeiten zu verstehen, vor- und zurückzublättern, Fehlantworten zu korrigieren, die Bearbeitung bestimmter Items auf einen späteren Zeitpunkt zu verschieben. Diese sind bei den derzeitigen Computersystemen nur eingeschränkt möglich.

Solche Neuerungen fordern den Probanden auf, sein Arbeits- und Reaktionsverhalten zu ändern (Jäger, R. S., 1990; Klieme & Stumpf, 1990).

Da die Computer-Variante von Papier-Bleistift-Tests die meist anzutreffende Umsetzungsform der Computerdiagnostik ist, wird das Problem der Äquivalenz im Rahmen eines Exkurses am Ende des Kapitels dargestellt.

18.2.1.2 Computer-Simulationstests

Computer-Simulationstests zeichnen sich durch die Nachbildung realer Situationen auf dem Computer aus. Sie zielen darauf ab, Aufschluß über Problemlösestrategien und das Entscheidungsverhalten von Probanden zu erhalten. Eine breite Anwendung finden sie im Bereich der Organisationspsychologie oder in der sogenannten „Management-Diagnostik“.

Der bekannteste Simulationstest ist der „MAILBOX-90“ (siehe Funke, J., 1993; Roest & Horn, 1990). Dieser stellt die computerisierte Form der im Rahmen von Assessment-Centers einsetzbaren Postkorbübung² dar und ist den sog. „Arbeitsproben“,³ zuzuordnen (Funke, J., 1993).

Weiterhin werden im Rahmen der „Management-Diagnostik“ Systemsimulationen konzipiert, die sich durch Komplexität, Dynamik, Intransparenz und Vernetztheit charakterisieren lassen (Dörner, 1989; Funke, U., 1993; Kluwe, 1995; Sonnenberg, 1993). Die Probanden werden dabei aufgefordert, die am Computer abgebildeten Systeme (z.B. ein fiktives Unternehmen, eine Stadt oder eine Insel) zu steuern und zu kontrollieren. Dabei können die gestellten Aufgaben je nach System differieren. Bei einigen Systemen sollen z.B. die Probanden Maßnahmen ergreifen, um bestimmte betriebliche Sollzustände zu erreichen. Bei anderen ist die Aufgabenstellung abstrakter definiert, z.B. ein fiktives Unternehmen möglichst gut zu leiten (Kluwe, 1995; Sonnenberg, 1993). Die kognitiven Anforderungen fallen dabei unterschiedlich aus.

Der komplexe, dynamische Charakter dieser Systemsimulationen soll die Ähnlichkeit zu realen Situationen herstellen. Deswegen besitzen sie - wie auch die Arbeitsproben - eine hohe „Augenschein-Validität“. Ungeprüft bleibt allerdings, ob die Realität oder der Arbeitsalltag so umfassend abgebildet wird, daß sich vom Testverhalten auf das tatsächliche Verhalten schließen läßt.

Als Versuch, die Realität genauer abzubilden, lassen sich die sogenannten „Videotests“ betrachten. Dabei wird die Fähigkeit des Computers genutzt, verschiedene Informationsquellen zu kombinieren: Ton-, Bild-, und Textinformationen. Durch sogenannte „true-to-life-situations“ können soziale Situationen „nicht nur statisch-photographisch in einem Printmedium sondern in beweg-

² Postkorb: Eine Übung im Assessment-Center, Einem Probanden werden Posteingänge, Notizen, Anfragen vorgelegt, die sich in Dringlichkeit, Komplexität und Bedeutsamkeit für die Firma erheblich unterscheiden. Der Proband muß alle Dokumente lesen und dann entscheiden, was geschehen soll. - Siehe Kapitel 23, S.491!

³ Arbeitsproben: (1) Im engeren Sinne: Aufgaben, die der Beobachtung konkreten Arbeitsverhaltens und des Handgeschicks dienen. Beispiel: Das Bild eines Kruges mit der Schere ausschneiden. - (2) Im weiteren Sinne: Aufgaben, welche die Anforderungen einer bestimmten Position möglichst realistisch abbilden, im Assessment-Center etwa die Übung des „Postkorbs“.

licher, lebensnaher Form als Filmsequenz“ dargestellt werden (Fricke, 1995, 579). Das interaktive Video oder Dialogvideo stellt ein neues, vielversprechendes Gebiet in der Computordiagnostik dar.

18.2.1.3 Adaptive Tests

Das Adaptive Testen umschreibt einen flexiblen Testprozeß, den die Probanden durch ihre Antworten mitbestimmen. Die Antwort auf ein Item entscheidet darüber, welches Item als nächstes zu bearbeiten ist. Auf diese Weise werden dem Probanden solche Items vorgelegt, die seinem Leistungsniveau angemessen sind.

Auswahl, Reihenfolge und Anzahl der Items, die zu bearbeiten sind, folgen bestimmten Algorithmen in Form fester Verzweigungsregeln oder komplizierter, psychometrischer Berechnungen. Abgeleitet sind solche Algorithmen aus probabilistischen Testmodellen. (Überblicke geben beispielsweise Hornke, 1977; Kisser, 1995; Weiss, 1982.)

Adaptive Tests, die nach festen Verzweigungsregeln konzipiert sind, zeichnen sich durch vorstrukturierte Itempools aus (Weiss, 1982).

Das nach den probabilistischen Testmodellen konzipierte Testverfahren stützt sich auf die Itemauswahl der simultan durchgeführten Algorithmen. Es sei hier grob der Ablauf solch einer Testung skizziert.

Ablauf einer adaptiven Testvorgabe

Alle Aufgaben werden aus einem großen *Itempool* gezogen, welcher folgende Voraussetzungen erfüllen soll:

- Der Pool enthält in ausreichendem Maße Items, die alle Schwierigkeitsstufen repräsentieren. Er ermöglicht somit eine Differenzierung auch in extremen Bereichen.
- Die Items sind „kalibriert“, das heißt: Sie haben sich empirisch als *homogen* erwiesen und ihre Parameter wurden geschätzt (z.B. Schwierigkeits-, Rate- oder Diskriminationsparameter) (Weiss, 1982).

Dieser Itempool kann als Item-Bank angelegt werden. Bei einer Testung werden Teilmengen bearbeitet, die, wie man annimmt, äquivalent sind. Die Äquivalenz kann empirisch mittels sogenannter „Equating Verfahren“ überprüft werden. (Equating Verfahren dienen dazu, die Eindimensionalität von Item-Teilmengen zu untersuchen.)

Die Testdurchführung beginnt in der Regel mit einem Item mittlerer Schwierigkeit. Löst der Proband das Item, ‚erhält‘ er ein schwierigeres. Löst er es nicht, folgt ein leichteres. - Nach dieser Anfangsphase erfolgt die Itemwahl nach Schätzung des Personenparameters. Geschätzt wird der Personenparame-

ter mittels maximum-likelihood-Schätzung⁴ oder mit Hilfe Bayesscher Formeln⁵. - Ausgewählt wird das Item, welches die größte Informationsfunktion besitzt. Die Informationsfunktion gibt an, „welchen Beitrag ein Item zur Einschränkung des Konfidenzintervalls des Personenparameters liefert“ (Kisser, 1995, 165). Es werden demnach stets Items ausgesucht, die den Personenparameter (das Fähigkeitsniveau) am fehlerfreiesten einschätzen. - Der Proband wird solange getestet, bis er ein bestimmtes Kriterium erreicht oder überschreitet. Als Abbruchkriterium dient in der Regel der Standardschätzfehler des Personenparameters.

Gegenüberstellung adaptiver und klassischer Testung

Aus dem geschilderten Testablauf wird ein wichtiger Unterschied zu der Testung „klassischer“ Art ersichtlich: Während der Itemsatz, die Reihenfolge der Items und die Testlänge bei der „klassischen“ Testung konstant bleiben, werden sie beim Adaptiven Testen durch die Antworten des Probanden bestimmt:

- In der klassischen Testung werden stichprobenabhängige Aussagen über unterschiedliche Merkmalsausprägungen verschiedener Personen gemacht, wenn sie sich dem gleichen Test unter möglichst gleichen Bedingungen unterzogen haben.
- In der Adaptiven Testung können Aussagen über interindividuelle Unterschiede gemacht werden, auch wenn die Probanden unterschiedliche Itemmengen bearbeitet haben (Kisser, 1995).

Die Vergleichbarkeit der Ergebnisse gewährleisten in diesem Fall die Itemparameter (z. B. Schwierigkeitsparameter), die weitgehend bekannt sein sollen. Die klassische Testtheorie sieht aber keine Trennung zwischen Item- und Personenparameter vor (Rasch-Modell: siehe S. 151).

Es wäre daher problematisch, sie als testtheoretische Basis für das Adaptive Testen heranzuziehen. Eine solche Trennung ist erst bei psychometrischen Modellen zulässig, die nach der sogenannten Item Response Theory“ (IRT) entwickelt wurden (Rettig & Hornke, 1990). Darunter fallen probabilistische Ansätze, wie das einparametrische Rasch-Modell (1980) oder die mehrparametrischen Birnbaum-Modelle (1968) (Klieme & Stumpf, 1990; Kubinger, 1995 a). In Anlehnung an das Rasch-Modell wurde beispielsweise ein Testalgorithmus für Adaptives Testen von Reckase (1974) entwickelt (Rettig & Hornke, 1990).

4 Maximum-likelihood-Methode: Ein statistisches Vorgehen, bei dem ein Parameter so geschätzt werden soll, daß der Standardfehler minimiert wird (Dorsch, 1994, 467).

5 Bayes-Theorem: Ein spezielles System der Statistik, das es gestattet, eine aposteriori-Wahrscheinlichkeit zu schätzen, d.h. die Wahrscheinlichkeit dafür, daß Ereignis B eintritt, nachdem Ereignis A bereits eingetreten ist (Dorsch, 1994, 87).

Bewertung des Adaptiven Testens

Durch adaptive Testalgorithmen wird eine Rationalisierung des diagnostischen Testprozesses angestrebt, die sich in *testökonomischen*, *meßqualitativen* und *motivationalen* Aspekten bemerkbar macht.

Unter *testökonomischen* Aspekten scheint es eindeutig, daß eine flexible Testprozedur zu einer Durchführungsökonomie führt: Indem die Testung durch das Antwortverhalten gesteuert wird, bleibt sie näher am Leistungsniveau des Probanden in bezug auf eine konkrete Fragestellung. Die „klassische“ Testung ist in dem Sinne unökonomischer, in dem sie Daten miterheben kann, die für die konkrete Fragestellung uninteressant bleiben, weil sie nicht informativ sind (Kisser, 1995).

Beispiel: Nehmen wir an, daß wir das Konstrukt „Raumvorstellung“ eines Erwachsenen erfassen möchten. Dafür stehen verschiedene Testverfahren zur Verfügung, z. B. Leistungs-Prüf-System (LPS: Horn, 1983), Intelligenz-Struktur-Test (IST 70: Amthauer 1970), Wilde-Intelligenz-Test (WIT Jäger A.O. & Althoff, 1984). Führen wir diese Testverfahren durch, so erhalten wir zusätzlich Angaben etwa über sprachliche Funktionen oder die Merkfähigkeit des Probanden, welche nicht unbedingt „informativ“ zur Erfassung der „Raumvorstellung“ sind.

Ein weiterer testökonomischer Aspekt betrifft die Testlänge. Beim Adaptiven Testen werden deutlich weniger Items gebraucht (zum Teil bis zu 50 Prozent), ohne daß die Testung an Reliabilität verliert. Das ergibt sich aus den bisher durchgeführten Studien (Wild, 1989). Aus dem Bestreben, mit dem Einsatz informativer Items die Testlänge zu verkürzen, resultiert nicht unbedingt eine Reduzierung der gesamten Testbearbeitungszeit (Kubinger, 1993). So fand Wild (1989) auf der Basis von Matrizen-Aufgaben, daß das Adaptive Testen zu längeren, z. T. verdoppelten Bearbeitungszeiten verleiten kann.

Unter *meßqualitativen* Aspekten ist der hohe Differenzierungsgrad des Adaptiven Testens hervorzuheben. Während konventionelle Tests eher unpräzise in den Extrembereichen differenzieren, ermöglichen adaptive Teststrategien eine Schätzung der Personenparameter auf allen Stufen des Fähigkeitskontinuums (Weiss, 1982). Dies resultiert aus dem zugrundeliegenden Itempool, dessen Items alle Schwierigkeitsniveaus abdecken sollen.

Die höhere Reliabilität des Adaptiven Testens stützt sich zudem auf Befunde, nach denen bei gleicher Testlänge adaptive Teststrategien zuverlässiger das Fähigkeitsniveau der Probanden schätzten als entsprechende klassische Tests (Klieme & Stumpf, 1990). Die Testlänge scheint aber eine Rolle zu spielen: McBride und Martin (1983) konnten diesen Befund nur bei Testlängen bis zu 20 Items bestätigen. Bei längeren Itemsätzen ergaben sich keine bedeutsamen Differenzen bezüglich der Reliabilität (Klieme & Stumpf, 1990).

Hinsichtlich der Validität adaptiver Teststrategien im Vergleich zu der klassischen Testung lassen sich die Ergebnisse verschiedener Studien folgendermaßen zusammenfassen:

- Bei kleinen Itemmengen (etwa 20 Items) zeigt sich das Adaptive Testen als überlegen gegenüber dem „klassischen“ Testen.
- Bei größeren Itemsätzen läßt sich eine Angleichung der Validitätswerte beobachten (Klieme & Stumpf, 1990).

Befürworter des Adaptiven Testens sehen in seinem Ablauf eine positive Wirkung auf die *Motivation* des Probanden: Aufgrund der testökonomischen Aspekte wird der Proband stets mit Aufgaben konfrontiert, die seinem Leistungsniveau entsprechen. Bei der „klassischen“ Testung wird jedem Probanden, unabhängig von seinem Fähigkeitsniveau, der gleiche Itemsatz vorgelegt. Dabei wirken zu leichte Items für einen leistungsstarken Probanden oder zu schwere Items für einen leistungsschwachen Probanden demotivierend oder frustrierend (Kisser, 1995; Kubinger, 1995).

Diese Erhöhung der Motivation unterliegt allerdings der Einschränkung, daß sie mit einer Erhöhung der Ängstlichkeit einhergehen kann. Ein wesentlicher Kritikpunkt zum Adaptiven Testen betrifft nämlich die Vernachlässigung von Rahmenbedingungen, wie einem dem Probanden angemessenen „Testklima“: dies kann Streß hervorrufen (Booth, 1995). Dazu trägt bei, daß dem Probanden am Anfang der Testung zu wenig leichte Items oder keine „Eisbrecher-Items“ vorgegeben werden, so daß die Aufwärmphase entfällt (Booth, 1995; Rettig & Horne, 1990).

Andererseits ist die Testängstlichkeit als ein mehrdimensionales Merkmal anzusehen, dessen Ausprägung nicht nur ein Ergebnis speziell des Adaptiven Testens ist, sondern auch von den verschiedenen Probanden-Typen abhängen kann.

Das Adaptive Testen löste bei den Diagnostikern einen großen Optimismus aus. Insbesondere durch den Einsatz des Computers im diagnostischen Prozeß konnten die Vorzüge des Adaptiven Testens zum Tragen kommen: maximaler Informationsgewinn unter testökonomischen Aspekten. Auch die Grundidee, durch einen antwortabhängigen Testablauf sensibler und präziser die „Individualität“ zu erfassen, könnte im Rahmen der behandlungsorientierten Diagnostik („Treatmentdiagnostik“) Verwendung finden: Eine individualisierte Messung erlaubt eine präzisere Erfassung von Veränderungen, wovon der diagnostisch/therapeutische Prozeß profitieren könnte (Jäger, R. S. & Krieger, 1994). Das computergestützte Adaptive Testen stellt zur Zeit eher ein „aktives“ Forschungsfeld dar. Seine Realisierung stieß auf einige Barrieren, welche zu einer kritischeren Gegenüberstellung führten.

- Zunächst sei *der große Aufwand* erwähnt, welcher mit der Konstruktion, Erprobung und Kalibrierung des erforderlichen Itempools verbunden ist. Für die Zusammenstellung einer solchen Itembank braucht man 100-200

Items. Die Gewährleistung der Homogenität bei solchen Itemsammlungen erweist sich als schwierig (Hageböck, 1994; Klieme & Stumpf, 1990; Wildgrube, 1990). Inhomogene Items - im Sinne eines bestimmten Testmodells - müssen ausgeschieden werden. Die verbleibenden Items müssen hinreichend das relevante Schwierigkeitskontinuum abdecken, um dem jeweiligen Leistungsniveau immer entsprechen zu können. Ansonsten kann es zu Unter- oder Überschätzungen der Personenparameter kommen (Kisser, 1995; Rettig & Hornke, 1990).

- Die *Auswirkung von Lernprozessen* auf die Ergebnisse bedarf weiterer Untersuchungen. Gemäß den Annahmen der probabilistischen Testtheorien ist die Lösungswahrscheinlichkeit eines Items unabhängig von seiner Position in der Reihenfolge, in der es vorgegeben wird. Empirische Befunde widerlegen allerdings diese Annahme. So ergab sich aus einer Studie von Wild (1989) folgendes: Ein Item, das gemäß seiner Kalibrierung mittelschwerig war, erwies sich als Startitem als deutlich zu schwierig. Somit sind Lern- und Übungeffekte beim Adaptiven Testen nicht auszuschließen.
- Neben der eher „inhaltlichen“ Problemstellung, sind die *Einschränkungen aufgrund der Computertechnologie* nicht zu übersehen. So weisen Jäger, R. S. und Krieger (1994) daraufhin, daß zur Zeit keine benutzerfreundlichen Systeme für die Implementierung adaptiver Strukturen zur Verfügung stehen.

18.2.2 Computer-Testsysteme, Computer-Testgeräte

Die Testsysteme, Testgeräte sind als eine Synergie von Hard- und Software zu verstehen mit dem Ziel, diagnostische Aufgaben auszuführen. Mittels sogenannter Peripheriegeräte und spezieller Programm-Module werden die Durchführung, die Registrierung der Probandenantwort und der zusätzlichen Daten, die Auswertung sowie die Verwaltung der Computertestverfahren ermöglicht. Die Ergebnisse werden in Form von Profilanalysen oder kurzen Interpretationen dargestellt. Eine kurze Rückmeldung oder „Gutachtenerstellung“ wird von manchen Testsystemen angeboten.

Das Angebot an Testsystemen oder Testgeräten ist groß und wird mit der Zeit immer unüberschaubarer. Zugunsten einer systematisierten Präsentation werden hier die Testsysteme folgendermaßen eingeteilt:

- *PC-Testsysteme, PC-Testgeräte*, welche auf einem Personal-Computer verwendbar sind,
- *große Testsystemanlagen*, die als Eigenentwicklungen großer Organisationen zu verstehen sind.

Hier werden exemplarisch vier PC-Testsysteme, PC-Testgeräte und drei eigens entwickelte Diagnosesysteme präsentiert.

Im deutschsprachigen Raum sind zur Zeit folgende *PC-Testsysteme*, *PC-Testgeräte* gängig (vgl. dazu Hageböck, 1994; Kisser, 1986; Klieme & Stumpf, 1990):

- Das *Computersystem „Test 3000“* der Firma ZAK. Es besteht aus einem Rechner, der als Testleiterplatz dient und bis zu acht Testplätzen. Die Testplätze sind über ein Interface mit dem Rechner verbunden. Sie bestehen aus einem Bildschirm und verschiedenen Geräten, wie beispielsweise Determinationsgerät, Tachistoskop usw. Das System eignet sich für sensumotorische Prüfungen.
- Die *PC- Version des „Wiener Testsystems“* der Firma Schuhfried. Wie funktioniert dieses System? Es liegt ein Zentralrechner als Testleiterplatz vor. Damit ist ein Probanden - Arbeitsplatz bestehend aus eigenem Bildschirm, (spezieller) Tastatur und Lichtgriffel verbunden. Zusätzlich können verschiedene Peripheriegeräte vom Zentralrechner angesteuert werden. Neben sensumotorischen Prüfungen bietet sich dieses System für Diagnosen im Bereich der Leistungs-, Persönlichkeits- und klinischen Psychologie hauptsächlich mittels Computer-Versionen von Papier-Bleistift-Tests an. Außerdem stellt es Autorensysteme zur Konstruktion von Computertests zur Verfügung.
- Das *„Rechnergestützte Psychodiagnostische System“ (RPS)* der Firma Horgreffe Apparatezentrum (Göttingen). Dieses System verfügt über eine spezielle Probanden-Tastatur und bietet ein weites Spektrum an Testverfahren, welche aufgabenspezifisch in sogenannte Programm-Module gruppiert sind. So enthält beispielsweise das „Modul“ LEILA (Leistungsdiagnostisches Labor) Verfahren zur Intelligenz- und Leistungsdiagnostik während das Modul PERSYS (Computerbasierte Persönlichkeitsdiagnostik) der Differentialdiagnostik von Persönlichkeit dient. Diese Programm-Module werden unter einer einheitlichen Benutzeroberfläche betrieben. Bei den Testverfahren geht es vornehmlich um computerisierte Formen gängiger Papier-Bleistift-Tests.
- Das *„Leipziger Testsystem“* besteht aus einem Personalcomputer (Testleiterplatz), der mit einem Kleinrechner (Probanden-Arbeitsplatz) sowie speziellen Peripheriegeräten verbunden ist. Er verfügt über ein umfangreiches Softwareangebot.

Aus der Präsentation der kommerziellen PC-Testsysteme wird ersichtlich, daß die Computerdiagnostik zur Zeit eher durch Computer-Versionen von Papier-Bleistift-Tests gekennzeichnet ist (S. 389). Bezüglich der Erprobung neuer Teststrategien (z. B. Adaptives Testen) zeigen sich große Organisationen offener als die kommerziellen Anbieter. Solche Organisationen verfügen über eigens entwickelte *Testsystemanlagen*. Nachstehend werden einige Beispiele angeführt:

- Das *System DELTA* (Dezentrales Testvorgabe- und Auswertungssystem im Psychologischen Dienst der Bundesanstalt für Arbeit) wird benutzt von

der Bundesanstalt für Arbeit (Hilke, 1993). Dabei geht es um ein vernetztes System, das einen Testleiter- und einen Sachbearbeiterplatz mit sechzehn Testplätzen verbindet. Die Testplätze können parallel angesteuert werden. Mit DELTA können unterschiedliche Testverfahren, darunter der schon bewährte „Berufswahltest“ (BWT, Bundesanstalt für Arbeit, 1991), durchgeführt, ausgewertet und die Ergebnisse in Form von Merkmalsprofilen und automatisierten Eignungsaussagen kommentiert werden. Adaptiv kann die Instruktion sowie die Testauswahl verlaufen. DELTA befindet sich zur Zeit in der Erprobungsphase und wird nur auf sechs Arbeitsämtern eingesetzt. DELTA wird im Rahmen der Berufseignungsdiagnostik zur Unterstützung der Entscheidungsprozesse von Ratsuchenden eingesetzt (Hilke, 1993).

- Die *Hard- und Softwareserie CAT* (Projekt: Computer Adaptives Testen) wird benutzt von der Bundeswehr und wurde entwickelt von den Firmen ZAK und Schuhfried (die Version CAT I), überarbeitet durch die Firma Dornier (die Versionen CAT II und CAT III). Dadurch können Computer-Versionen von Papier-Bleistift Intelligenz- und Leistungstests vorgegeben werden. Ab der Version CAT II wird die computergestützte Form der Basis Eingangs-Testbatterie (Eignungs- und Verwendungs-Testbatterie: EVT) dargeboten. Weiterhin können Spezialuntersuchungen (z. B. für Beamtenanwärter, Panzerfahrer, Fremdsprachendienst, usw.) oder Simulationstests für Luftfahrzeugführeranwärter (Piloten, Kampfbeobachter) durchgeführt werden, Inzwischen liegt die dritte Version der CAT-0Anlage (CAT III) vor. Sie verfügt über einen Adaptiven Algorithmus, welcher zunächst nur experimentell eingesetzt werden soll (Hageböck, 1994; Rauch, Weber & Wildgrube, 1993; Wildgrube, 1990). Das CAT-System dient Platzierungsentscheidungen und der Beratung oder Zuweisung zu verschiedenen Laufbahnen und Tätigkeitsbereichen bei jungen Wehrpflichtigen und Langdienern.
- Das *Computestgerät ART-90* (Act und React Testsystem), wird benutzt vom Österreichischen Kuratorium für Verkehrssicherheit und wurde entwickelt in Zusammenarbeit mit der Firma Schuhfried (Bukasa, Kisser & Wenninger, 1990). Es besteht aus einem Computer, spezieller Hardware zur Dateneingabe (Lichtgriffel, Hand-, und Fußtasten) und Peripherie-Geräten (wie z. B. Probanden-Bildschirm, Cognitrone, Wiener Determinationsgerät II, Drehpotentiometer). Diese Testgeräte ermöglichen es, spezielle Leistungsverfahren (z. B. zur Messung des Reaktionsvermögens, der Sensumotorik oder der visuellen Wahrnehmung) durchzuführen und auszuwerten. Für das gesamte System liegen neue Normwerttabellen und Validierungsstudien vor (Bukasa, Kisser & Wenninger, 1990; Hageböck, 1994). Das System dient der Messung der Fahreignung.

Neben den genannten umfangreichen Systemen liegen einfache Computerprogramme vor, wie beispielsweise für qualitative Analysen das Programmsystem AQUAD (Analyse Qualitativer Daten: Huber, G.L., 1989) oder zur Auswer-

tung und Interpretation spezieller Verfahren das Rorschach-Interpretationssystem.

18.2.3 Computer-Interpretationssysteme

Bei den Interpretationssystemen geht es um Diagnosesysteme, welche die Urteilsbildung des erfahrenen Diagnostikers zu simulieren versuchen. Dafür stehen zwei Modellansätze zur Verfügung (Hageböck, 1994):

Der *Lineare Modellansatz* („bootstrapping“): Empirische Befunde unterstützen die These, daß die Urteils- und Entscheidungsfindung eines erfahrenen Diagnostikers durch einfache lineare (Regressions)-Modelle adäquat abbildbar sind.

Der *Prozeß-Modellansatz* („Process-Tracing“⁶): Die Abbildung der (klinischen) Vorgehensweise des Diagnostikers resultiert in diesem Fall aus der Analyse seiner explizierten Denk- und Entscheidungsprotokolle. Dementsprechend kann sie in Form von Flußdiagrammen abgebildet werden, welche in Computerprogramme umsetzbar sind.

Beide Modelle unterscheiden sich nur im Allgemeinheitsgrad und können bei Interpretations- oder Expertensystemen komplementär eingesetzt werden.

Die einfachsten Interpretationssysteme (auch als „deskriptiv“ zu bezeichnen) beinhalten eine kurze Erläuterung des Tests und seiner Konstrukte und abschließend stufen sie die erbrachte Leistung in den entsprechenden Ausprägungsbereich ein (als überdurchschnittlich, durchschnittlich oder unterdurchschnittlich).

Die zwei Modellansätze des Diagnostikers finden Anwendung in Interpretationssystemen, die eine weitere Verarbeitung der Ergebnisse erlauben. Ein Beispiel dafür stellen die verschiedenen MMPI-Interpretationssysteme⁶ dar. Neben der Interpretation der einzelnen Skalen ermöglichen sie die Integration von Einzelergebnissen in übergeordnete Befunde im Sinne von kurzen, automatisierten Gutachten.

Schließlich sind die Diagnosesysteme zu erwähnen, die eine entscheidungsunterstützende Funktion im diagnostischen Prozeß übernehmen können. Sie erlauben die Verarbeitung vorhandener Testergebnisse auch hinsichtlich bestimmter Fragestellungen, die vom Diagnostiker für den Einzelfall vorgegeben werden. Als Beispiel sei genannt das Computersystem DIASYS (Computer-gesteuertes diagnostisches System auf normativer Grundlage: Hageböck, 1994). Zur Beurteilung von Schul- oder Lernschwierigkeiten und Verhaltensauffälligkeiten im Rahmen der schulpädagogischen Einzelfallhilfe bietet das

⁶ MMPI: Minnesota Multiphasic Personality Inventory. Ein mehrdimensionaler Fragebogen. Siehe: Hathaway & McKinley (1951).

System eine Liste von Hypothesen zur Auswahl an, die auf der Basis von zugrundeliegenden Test-Scores auf Gültigkeit (Akzeptanz oder Ablehnung) überprüft werden. Die Ergebnisse können im Rahmen eines Gutachtens verwendet werden.

Die automatisierte Interpretation von Ergebnissen kann zur Standardisierung und Ökonomisierung der Testung beitragen. Man zielt darauf ab, die Störanfälligkeit klinischer Urteile (z.B. durch eine Übermüdung des Experten) auszuschließen. Befürworter einer computergestützten Interpretation verzeichnen als Ergebnis vieler Untersuchungen, daß die statistische Urteilsbildung der klinischen hinsichtlich ihrer Zuverlässigkeit überlegen sei (Hofer & Green, 1985). (Einen Überblick bietet Hageböck, 1994.) Eine hohe Reliabilität bringt allerdings nicht gleichzeitig eine hohe Validität mit sich. Die Validität solcher Systeme stellt den neuralgischen Punkt dar und wurde zum Mittelpunkt kontroverser Diskussionen.

Matarazzo (1983) weist auf die Gefahr des Mißbrauchs solcher Systeme hin: Automatisierte Interpretationen besitzen eine „Schein“-Objektivität und können deshalb nicht unkritisch akzeptiert werden. Das Problem wird verschärft, wenn Laien sie erhalten, die ihre Begrenzungen nicht kennen. Bis ihre Validität von mehreren Untersuchungen bewiesen wird, sollten solche Systeme nur von qualifizierten Personen angewandt werden.

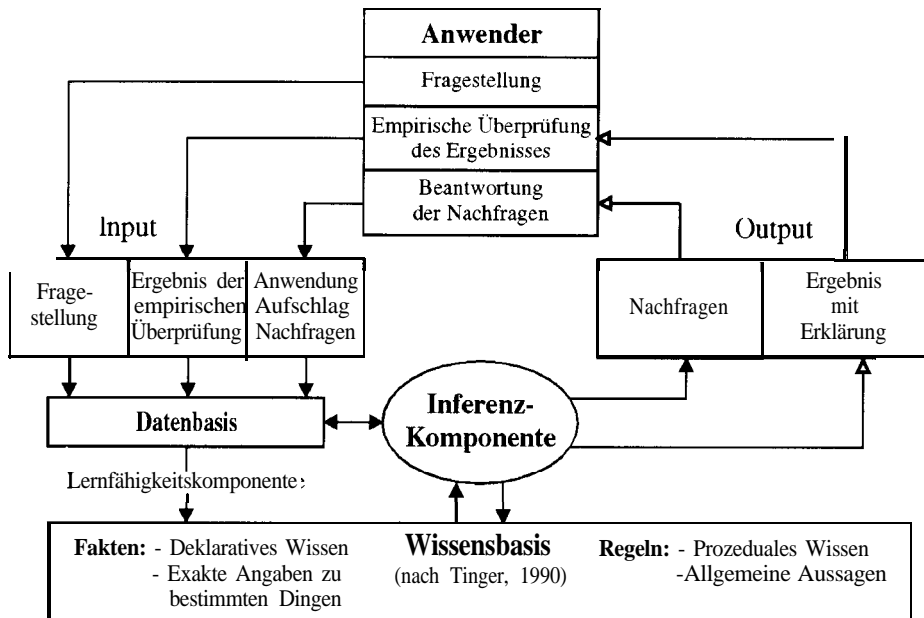
Die Überprüfung der Validität erweist sich allerdings als schwierig: zum einen fehlen die geeigneten Methoden und zum anderen liegen uneinheitliche Untersuchungsergebnisse vor (Hageböck, 1994). Dabei spielt die Konzipierung und Verfügung dieser Systeme eine Rolle: Während die Grundidee ist, den Kenntnisstand von Experten möglichst vollständig zu simulieren, werden stattdessen die Kenntnisse des Systementwicklers wiedergegeben (Hageböck, 1994). Weiterhin mangelt es bei den meisten solcher Systeme an Transparenz: Die einzelnen Schritte bis zur Interpretation der Ergebnisse werden von den Herstellern nicht mitgeteilt (Klieme & Stumpf, 1990). Daher ist eine Überprüfung automatisierter Interpretationen von qualifizierten Personen notwendig. Das Testkuratorium empfiehlt sogar das Heranziehen eines „relevanten Außenkriteriums“ (1986, 164).

18.2.4 Computer-Expertensysteme

Die Expertensysteme wurden bei den Forschungen zur Künstlichen Intelligenz entwickelt. Es handelt sich um Diagnosesysteme, die das Fachwissen von Experten und ihre Problemlösetechniken möglichst umfassend abbilden sollen - soweit sie sich auf Fragestellungen in einem spezifischen Anwendungsbereich beziehen (Hageböck, 1994; Tinger, 1990).

Die Konzipierung solcher Systeme setzt ein umfangreiches Wissen auf dem betreffenden Fachgebiet voraus, welches auch in ein Computersystem umsetz-

Kasten 18-2:
Schema eines Expertensystems
 Siehe laufenden Text!



bar sein muß (Hageböck, 1994). Sämtliche Kenntnisse auf dem Fachgebiet, strukturiert in Form von Verarbeitungsregeln, stellen dann die *Wissensbasis* des Systems dar. Die Wissensverarbeitung erfolgt dann vor dem Hintergrund bestimmter Fragestellungen und Informationen, die der Benutzer über ein Dialogfeld (Input) in das System eingegeben hat. Diese werden in der *Datenbasis* abgelegt. Über eine sogenannte „*Inferenzkomponente*“, welche als das „Problemlösungs-Programm“ verstanden werden kann (Tinger, 1990, 206), werden die in der Wissensbasis enthaltenen Verarbeitungsregeln mit den eingegangenen Fragestellungen und Informationen so kombiniert, daß daraus logische Schlüsse zur Beantwortung dieser Fragestellungen gezogen werden können. Der ganze Ablauf erfolgt nach dem heuristischen Prinzip: Für jede Frage stehen verschiedene Problemlösungen zur Verfügung. Während der Wissensverarbeitung können weitere Informationen vom System abgerufen oder vom Anwender als Antwort auf Nachfragen erhalten werden. Diese Informationen werden zunächst überprüft und dann zu einem logischen Schluß geführt (Output). Die einzelnen Problemlösungsschritte sind durch eine „*Erklärungskomponente*“ begründet, so daß man den Entscheidungsprozeß verfolgen und überprüfen kann. Manche Expertensysteme verfügen über eine „*Lernfähigkeitskomponente*“, die eine Modifizierung und Erweiterung der Wissensbasis erlauben. Dies kann durch neu eingegangener Fälle oder der Verwertung der Ergebnisse erfolgen. (Ein Beispiel zur Wissensrepräsentation und Wissensverarbeitung mittels Expertensysteme bietet Ueckert, 1995.)

Eine schematische Darstellung des Expertensystems ist in Kasten 18-2 wiedergegeben.

In der psychologischen Praxis finden solche Expertensysteme im klinischen Bereich eine Anwendung. Ihr Einsatz betrifft dann Entscheidungen, die eine automatisierte Zuordnung von Fällen in vorgegebene Klassifikationssysteme ermöglichen. Ein Beispiel dafür stellt das Expertensystem DSM-III-X (Diagnostic and Statistical Manual of Mental Disorders: Langner, 1988) dar.

Expertensysteme werden zur Zeit eher für Forschungszwecke verwendet. Praktisch angewendet werden nach Schätzungen weniger als zwei Prozent der bisher entwickelten Systeme (Hageböck, 1994).

18.3 Einsatzfelder für eine computergestützte Diagnostik

Computerdiagnostik wird auf verschiedenen Gebieten eingesetzt:

- **Im organisations- und betriebspsychologischen Bereich** dienen Computerprogramme der Personalauswahl und -auslese. Es werden sowohl computerisierte Persönlichkeits- und Einstellungstests als auch computergestützte Arbeitsproben (z.B. „MAILBOX-90“), Simulationstests und Planspiele eingesetzt. Computer werden weiterhin zur Auswertung von Assessment-Center Ergebnissen, zur Bewertung von Bewerberunterlagen, zur Ermittlung von Anforderungsprofilen oder zur Integration von Befunden genutzt (Kluwe, 1995; Sonnenberg, 1993).
- **Im klinischen Bereich** können Computer-Versionen von Papier-Bleistift-Tests als Screening-Verfahren eingesetzt werden. Darüber hinaus können Probanden ihre biographischen Daten oder standardisierten Interviews direkt in den Computer eingeben. Es stehen Auswertungsprogramme für projektive Verfahren zur Verfügung (z.B. zum Rarschach-Test). Im klinisch-psychiatrischen Bereich wird - allerdings eingeschränkt - vom Klassifikationssystem DSM-III-X (Langner, 1988) Gebrauch gemacht. Im Rehabilitationsbereich werden Computertests zum Training von Leistungsfähigkeiten (z. B. zum Gedächtnistraining) oder zur Sensumotorik eingesetzt (z. B. REHACOM: kognitive Rehabilitation).
- **Im pädagogisch-psychologischen Bereich** werden Computer z.B. zur Herstellung von Lern- und Übungsprogrammen, zur Intelligenzmessung oder zur Überprüfung von Schul- und Lernschwierigkeiten (z.B. DIASYS) verwendet. Es liegt ein Item-Bank-System für Prüfungen an den Industrie- und Handelskammern vor. Weiterhin stehen spezielle Adaptive Computertests zur Verfügung, z. B. ADAFI (Adaptiver Figurenfolgen-Lerntest: Guthke, 1989).
- **In Großorganisationen** werden zu Zwecken der Eignungsdiagnostik eigene Testbatterien entwickelt, die auch computerunterstützt dargeboten werden.

18.4 Zusammenfassung zu Kapitel 18

Die Nutzung des Computers in der psychologischen Diagnostik reicht von der Ausführung einzelner Aufgaben bis hin zur Planung und Entwicklung komplexer Diagnosesysteme, welche die Vorgehensweise eines Experten zu simulieren versuchen.

Auf der Testebene bietet der Computer Möglichkeiten, Tests ökonomischer zu entwickeln (durch die automatische Itemgenerierung und die Anlage von Item-Banken). Darüber hinaus können während der Durchführung zusätzliche Daten erhoben, sowie neben der Datenverwertung die Ergebnisse interpretiert und in übergeordnete Befunde integriert werden.

Computerdiagnostik wird durch die Planung und Entwicklung von Computersystemen umgesetzt. Sie wurden in drei Gruppen eingeteilt: *Computertests*, *Computer-Interpretationssysteme* und *Computer-Expertensysteme*.

Drei Varianten von *Computertests* wurden diskutiert:

- Computer-Versionen von Papier-Bleistift-Tests*: Sie stellen die häufigste Erscheinungsform computerunterstützter Diagnostik dar. Die Computer-Vorgabe trägt zu einer höheren Standardisierung und Zeitersparnis der Testung bei. Allerdings muß eine Überprüfung der Äquivalenz zwischen den beiden Darbietungsmodi vorgenommen werden.
- *Simulationstests*: Sie lassen sich durch die Nachbildung realer Situationen charakterisieren, mit dem Ziel, das Problemlöse- und Entscheidungsverhalten zu erfassen. Aufgrund ihres realitätsnahen Charakters wird ihnen eine hohe „Augenschein-Validität“ beigemessen. Inwieweit sich die komplexe Realität in solchen Tests abbilden läßt, bleibt der Kritikpunkt der Simulationstests. Eine Lösungsmöglichkeit des Problems könnten „Videotests“ anbieten, durch welche eine reale Situation nicht nur statisch, sondern als eine „Filmsequenz“ abgebildet werden kann.
- *Adaptive Tests*: Sie ermöglichen einen flexiblen Testablauf, welcher durch das Antwortverhalten der Probanden bestimmt wird. Sie bauen auf probabilistischen Testmodellen auf. Den vielversprechenden Vorteilen des Adaptiven Testens (Durchführungsökonomie; verkürzte Testlänge; dem jeweiligen Leistungsniveau angemessene Testung; höher Leistungs-Differenzierungsgrad auch in den extremen Bereichen; Steigerung der Motivation der Probanden) stehen gewisse Einschränkungen entgegen (großer Aufwand bei der Konstruktion, Erprobung und Kalibrierung des erforderlichen Itempools; Gewährleistung zufriedenstellender Homogenität; trotz verkürzter Testlänge z.T. verdoppelter Bearbeitungszeit; Erhöhung der Testängstlichkeit; Lern- und Übungseffekte).

Die *Computer-Testsysteme*, *Computer-Testgeräte* sind als eine Synergie von Hard- und Software zu verstehen, mit dem Ziel diagnostische Aufgaben aus-

zuführen (z. B. Durchführung von Testverfahren, Auswertung und Interpretation der Daten, Verwaltung großer Datenmengen). Es wurden präsentiert:

- Vier kommerzielle *PC-Testsysteme*, *PC-Testgeräte*. Mit Ausnahme der sensumotorischen Tests, bieten diese Systeme hauptsächlich Computer-Versionen von Papier-Bleistift-Tests an.
- Drei große *Testsystemanlagen*. Sie wurden von großen Organisationen zu eigenen diagnostischen Zwecken entwickelt. Im Gegensatz zu den kommerziellen Anbietern zeigen sie sich zugänglicher bezüglich der Erprobung neuer Teststrategien (z.B. Adaptives Testen).

Bei der Entwicklung von Computer-Interpretationssystemen wird versucht, den erfahrenen Diagnostiker zu simulieren. Dafür stehen der *Lineare Modellansatz* und der *Prozeß-Modellansatz* zur Verfügung. Die Interpretation der Ergebnisse reicht von der einfachen Einstufung der erbrachten Leistung in dem entsprechenden Ausprägungsbereich, über ihre Integration in übergeordnete Befunde bis hin zur gezielten Datenverarbeitung hinsichtlich bestimmter Fragestellungen. Der erwarteten hohen Standardisierung und Ökonomisierung der Testung steht vor allem das Problem der Validität solcher Systeme gegenüber.

Die Simulation des erfahrenen Diagnostikers in seiner Urteilsbildung erreicht ihre Grenzen bei den sogenannten Computer-Expertensystemen. Ihre Struktur erlaubt die Vorgehensweise nach einem heuristischen Prinzip. Das heißt, im Gegensatz zu den anderen Computersystemen, die mit festgelegten Algorithmen arbeiten, verfügen Expertensysteme über verschiedene Problemlösungswege. Die Auswahl der Problemlösung erfolgt nach der gezielten Auswertung des schon vorhandenen Wissens, eventuell kombiniert mit den Antworten, die der Anwender auf Nachfragen des Systems gibt. Die Anwendung von Expertensystemen befindet sich in einer Erprobungsphase.

18.5 Kontrollaufgaben zu Kapitel 18

- Automatisierte Itemgenerierung.
- Item-Banken.
- Äquivalenz zwischen Papier-Bleistift- und Computer-Versionen: psychometrische, subjektiv erlebte, Testgüte.
- Systemsimulationen, „Videotests“.
- Adaptives Testen: Begriffserklärung, Ablauf, Vor- und Nachteile.
- Testsysteme, Testgeräte: Varianten, Beispiele.
- Computer-Interpretationssysteme: Modellansätze, Kritikpunkte.
- Computer-Expertensysteme: Struktur, diagnostische Relevanz.

Exkurs: Zur Äquivalenz zwischen Papier-Bleistift- Tests und ihren Computer-Versionen

In Anlehnung an Honaker (1988) läßt sich die Äquivalenz zwischen der Papier-und-Bleistift-Form und der Computer-Version desselben Verfahrens in eine psychometrische Äquivalenz und in eine subjektiv erlebte Äquivalenz aufgliedern. Bei der psychometrischen Äquivalenz unterscheidet Marco (1981) zwischen Mittelwertsverschiebungen, Änderung der Metrik (Anderung wichtiger Verteilungskennwerte, vor allem der Standardabweichung) und Änderung der Reihenfolge der Items (Jäger, R. S., 1990; Klieme & Stumpf, 1990).

Im folgenden wird die Äquivalenz unter zwei Gesichtspunkten behandelt:

- Äquivalenz auf der Meßebeane (A),
- Äquivalenz in der Akzeptanz der Probanden (B).

(A) Äquivalenz auf der Meßebeane

Ob Papier-Bleistift-Tests und ihre Computer-Versionen äquivalent auf der Meßebeane sind, wird unter zwei Gesichtspunkten geprüft:

- Psychometrische Äquivalenz (A-a),
- Äquivalenz der Testgute (A-b).

(A-a) Psychometrische Äquivalenz

Die Änderung der Metrik ist aus der Sicht der psychologischen Messung als die wichtigste anzusehen, da sie Unterschiede bei grundsätzlichen Verteilungswerten der Daten nach sich zieht. Änderungen in den Rangordnungen sind ausschlaggebend aus der Sicht der Probanden, da sie eine eventuelle Selektion des betreffenden Probanden in Frage stellten (Jäger, 1990).

Gemäß den bisher durchgeführten Äquivalenzstudien fallen die Resultate bei Leistungs- und Persönlichkeitstests unterschiedlich aus.

Es liegen viele Äquivalenzstudien **zu Leistungstests** vor, besonders zu den Standard-Progressive Matrices (SPM: Raven, 1971)‘.

- Diese Studien weisen auf Mittelwertsunterschiede zwischen den beiden Darbietungsformen hin. In einigen Untersuchungen erwiesen sich die Mittelwertsverschiebungen als unbedeutend. (Collins & Odell, 1986; Rock & Nolen, 1982). Sowohl Neubauer, Urban und Malle (1991) als auch Schwenkmezger und Hank (1993) fuhren dieses Ergebnis allerdings auf methodische Mängel zurück.
- Andere Studien stellten bedeutsame Mittelwertsunterschiede fest (Beaumont & French, 1987; Hilke, 1993; Kubinger & Farkas, 1991; Neubauer, Urban & Malle, 1991). Bei signifikant erwiesenen Mittelwertsunterschieden ergaben sich schlechtere Leistungen in der Computer-Form (Mittelwerte von Papier-Bleistift-Tests lagen signifikant höher als die der Computer-Versionen).

Man versucht solche Ergebnisse zu interpretieren wie folgt: Die Aufgaben- oder Itempräsentation am Bildschirm verleitet zu einer vorschnellen - und deshalb nicht sorgfältigen - Bearbeitung der Aufgaben. Es liegt die Vermutung nahe, daß der Computer einen „streßbevozierenden Charakter“ besitzt (Kubinger, 1993; Neubauer, Urban & Malle, 1991). Dabei spielt das Auflösungsvermögen des Computerbildschirms auch eine Rolle. Beaumont und French (1987) erklären das frühere Aufgeben der Bearbeitung mancher Aufgaben durch ihre mangelnde Graphik-Darstellung am Computer.

- Weiterhin merken Klieme und Stumpf (1990) an, daß eine Abweichung der computerisierten Itempräsentation vom Testheft der Papier-und-Bleistift-Version mit Schwierigkeitsunterschieden einhergeht. „Ein solcher Fall liegt etwa vor, wem-i komplexe Zuordnungsaufgaben, die im Testheft auf eine Seite passen, auf dem Bildschirm nur in Teilen sukzessiv dargestellt werden können“ (Klieme & Stumpf, 1990, 21).
- Die vorliegenden Mittelwertsunterschiede bringen allerdings nicht unbedingt bedeutsame Differenzen in den Streuungskennwerten mit sich. Die im Rahmen der Leistungsdiagnostik in der Deutschen Bundeswehr durchgeführten Äquivalenzprüfungen stellten trotz Mittelwertsunterschieden zwischen den Papier-Bleistift und den Computer-Versionen ähnliche Standardabweichungen fest (Wildgrube, 1990). So weisen Klieme und Stumpf (1990) auf eine gewisse Unabhängigkeit der Varianz der Testwerte vom Darbietungsmedium hin.

Aufgrund der Ergebnisse ist die Äquivalenz zwischen Papier-Bleistift-Tests und ihren Computer-Versionen (als Parallel-Tests) nicht als selbstverständlich zu betrachten. Solange die Äquivalenz nicht nachgewiesen ist, müssen neue Normen ermittelt werden (Bukasa et al., 1989; Klieme & Stumpf, 1990; Kubinger, 1993; Tanzer, 1987).

Bezüglich der Ermittlung neuer Normen schlägt Wildgrube (1990) vor, die Tageszeit der Testung mitzubersichtigen. In den Untersuchungen bei der Deutschen Bundeswehr traten nämlich bedeutsame Schwankungen und Unterschiede in den Leistungen auf, abhängig von der Tageszeit der Bearbeitung.

Als Folge könnte man sogar unterschiedliche Normen je nach Tageszeit der Testbearbeitung bilden (Wildgrube, 1990).

Die Äquivalenz von **Persönlichkeitstests** wurde am häufigsten mit dem MMPI geprüft. Schwenkmezger und Hank (1993) berichten in Anlehnung an Honaker (1988) daß bei den Skalen zwar Mittelwertsdifferenzen auftraten, aber nicht signifikant waren. Auch die Streuungs- und Verteilungsmaße wichen nicht signifikant voneinander ab. Im allgemeinen sprachen die hohen Korrelationen für eine Äquivalenz zwischen der Papier- und der Computer-Version des MMPI. In ihrer Studie fanden Bader, Hofmann und Kubinger (1993) keine signifikanten Mittelwerts- und Streuungsunterschiede zwischen der Papier- und-Bleistift-Version und der Computer-Form des Gießen Tests. Beide Versionen erwiesen sich somit als äquivalent. Wilson, Genco und Yager (1985) unterstützen mit ihrer Studie (zur Attitude Battery) auch die These der Äquivalenz zwischen Papier-Bleistift- und Computer-Formen von Persönlichkeits-tests. Klieme und Stumpf (1990) berichten allerdings über eine Minderzahl von Studien, bei denen statistisch signifikante Mittelwertsunterschiede nachgewiesen wurden. Beaumont und French (1987) fanden für die Persönlichkeitsskalen von Eysenck bei der Computer-Form höhere Neurotizismus-Werte als bei der Papier- und-Bleistift-Version. Schwenkmezger und Hank (1993) konnten eine Äquivalenz zwischen beiden Darbietungsmodi, zumindest bei situationsspezifischen Meßinstrumenten, nicht erkennen.

Aufgrund der Befunde erweist es sich als unzulässig, die Ergebnisse einer Studie auf alle Persönlichkeitsfragebogen zu generalisieren. Im Unterschied zu Leistungstests zeichnet sich aber deutlich die Tendenz ab: Die psychometrische Äquivalenz bei Persönlichkeitstests liegt höher als bei Leistungstests.

Schließlich blieb in den beiden Darbietungsformen die Rangordnung der Probanden erhalten (Klieme & Stumpf, 1990).

(A-b) Äquivalenz der Testgüte

Die Testgüte wird nach den üblichen Haupt- und Zusatzkriterien beurteilt:

- Hauptkriterien sind: *Objektivität, Reliabilität und Validität.*
- Zusatzkriterien sind: *Normiertheit, Ökonomie, Nützlichkeit, Vergleichbarkeit.*

Hauptkriterien

Bezüglich der drei Arten von *Objektivität* ist anzumerken:

- Die Durchführungsobjektivität ist bei der Computer-Version höher aufgrund des Ausschlusses von Testleitereffekten. Ob irgendwelche „Computer-Effekte“ eintreten, wird wahrscheinlich Thema zukünftiger Forschungen sein.

- Die Auswertungsobjektivität ist bei den Computer-Versionen selbstverständlich gegeben, da für den Computer die Auswertung insbesondere gebundener Items, eine Routine-Arbeit darstellt.
- Die Interpretationsobjektivität ist sowohl bei den Papier-Bleistift-Tests als auch bei deren Computer-Formen gesichert.

Was die *Reliabilität* anbetrifft, so ist anzumerken:

- Bei Leistungstests fanden Sacher und Fletcher (1978) sowie Neubauer, Urban und Malle (1991) ähnliche Reliabilitätskoeffizienten bei Papier-Bleistift- und Computer-Formen. Jäger, R. S. (1990) beschränkt den Geltungsbereich dieses Ergebnisses auf die Fälle, bei denen durch das Format und andere Ähnlichkeiten der Itempräsentation eine äußere Äquivalenz beider Testverfahren gegeben ist. Bei Persönlichkeitstests wird über eine gute Entsprechung der Reliabilitäten von Papier-Bleistift- und ihre Computer-Versionen berichtet (Klieme & Stumpf, 1990).

Zur *Validität* sei auf die Richtlinien des Testkuratoriums (1986) verwiesen:

- Nach diesen dürfen „Ergebnisse von Validitätsstudien nicht ungeprüft von der manuellen auf die EDV-gestützte Testung übertragen werden“ (S. 164). In der Praxis wird gegen diese Regel häufig verstoßen. Die psychologische Aussagekraft der Ergebnisse wird insbesondere durch die ungeprüfte Verwendung der Papier-Bleistift Normen auf die Computer-Version beeinträchtigt. Ansonsten wird in bezug auf Konzentrationstests und Arbeitsproben der Computer-Vorgabe „Augenschein“-Validität oder logische Validität zugeschrieben (Kubinger, 1993; Sonnenberg, 1993).

Zusatzkriterien

Das Problem der *Normierung* ergibt sich, wenn die Normtabellen von Papier-Bleistift-Tests auf die Computer-Version ungeprüft übertragen werden. Das Problem scheint bei Persönlichkeitstests nicht so groß zu sein. Bader, Hofmann und Kubinger (1993) fanden, „daß Persönlichkeitsfragebögen in bezug auf Normen-Verschiebungen tatsächlich wenig anfällig sind“ (Kubinger, 1993, 132). Trotz allem wird „im Regelfall eine neue Testeichung erforderlich sein“ (Testkuratorium, 1986, 164).

Nur wenige Testverfahren verfügen über computerbezogene Normen. Soweit sie zur Verfügung stehen genügen sie aber nicht dem Kriterium der Repräsentativität (vgl. Hageböck, 1994). Werden Computer-Normen erhoben, dann ist folgender Vorteil zu verzeichnen: Anhand von Norm-Banken können die Normen laufend aktualisiert werden.

Der Computer hat durch die Automatisierung von Teilaufgaben und der Individualisierung der Testung zur *Ökonomie* des testdiagnostischen Prozesses beigetragen. Einzelne Aufgaben, wie Itemkonstruktion (Item-Banken), Revision eines Tests, Testauswertung, verbale und graphische Interpretation der Ergebnisse und Testinstruktion fallen mit Computer einfacher als bei Papier-Blei-

Stift-Tests und können zu einer Verkürzung der gesamten Durchführungszeit führen.

Lienert und Raatz (1994) messen einem Test hohe *Nützlichkeit* bei, „wenn er in seiner Funktion durch keinen anderen vertreten werden kann“ (S. 13). In diesem Sinne erweist sich die bloße Übertragung von Papier-Bleistift-Tests auf den Computer nicht als nützlicher (Kubinger, 1993).

In bezug auf das Kriterium der *Vergleichbarkeit* gilt: Die Computer-Version eines Papier-Bleistift-Tests könnte mehr Vergleichsparameter anbieten. Darunter fallen beispielsweise die vom Computer automatisch registrierten Latenzzeiten, Fehlerkorrekturen, Wartezeiten usw.

(B) Äquivalenz in der Akzeptanz der Probanden

Den Vorteilen einer computergestützten Testung wird ein großer Nachteil gegenübergestellt: Der Computer steigert bei den Probanden die Test-Ängstlichkeit. Der Forschungsstand bietet diesbezüglich kein einheitliches Bild. Die Ergebnisse fallen verschieden für Leistungs- und Persönlichkeitstests aus.

Bei **Leistungstests** sind im deutschsprachigen Raum Akzeptanzbefragungen durchgeführt worden

- (a) bei der Deutschen Bundeswehr (Wildgrube, 1990) und
- (b) bei der Bundesanstalt für Arbeit (Hilke, 1993).

Zu (a): Die Akzeptanzbefragung bei der Deutschen Bundeswehr stammt aus den Jahren, 1986/87. Eine Stichprobe, repräsentativ bezüglich der Schulbildung, bewertete anhand eines Fragebogens die Bearbeitung desselben Tests

- in der Computer-Version (N = 884)
- und
- in der Papier-Bleistift-Form (N = 613).

Die Ergebnisse sprechen für eine hohe Akzeptanz der computergestützten Testung, welche im Vergleich zu der Papier-Bleistift-Testung als positiver eingestuft wurde.

Die Computer-Vorgabe wurde darüber hinaus als benutzerfreundlicher empfunden. Betont wurde, daß Korrekturen am Bildschirm gegebenenfalls einfacher vorzunehmen sind als mit Bleistift und Radiergummi.

Dem Computer als Medium wurden Eigenschaften wie „modern“, „ruhig“, „klar“ oder „sachlich“ zugeschrieben. Vorerfahrungen spielen allerdings eine Rolle: „Das computergestützte Testen wird eher von denen abgelehnt, die sich nichts darunter vorstellen können. Vorerfahrungen mit Heimcomputern und spätestens die Teilnahme an einem bildschirmgestützten Testverfahren führen

zu einer deutlichen Einstellungsänderung zu Gunsten dieser neuen Technik“ (Wildgrube, 1990, 145).

Zu (b): Die Akzeptanzbefragung bei der Bundesanstalt für Arbeit fand 1993 statt. 1500 Ratsuchende wurden nach ihrer Meinung zu beiden Darbietungsformen befragt (Männer und Frauen, Altersspanne: 13-48 Jahre). Insgesamt lautete das Urteil: Die Computer-Version ist angenehmer als die Papier-Bleistift-Version. Jugendliche stuften die Computer-Variante besonders hoch ein.

Ein weiterer interessanter Befund besagte: Jugendliche, die Sonderschulen für Lernbehinderte besucht hatten, machten sich mit der Maus als Eingabemedium genauso schnell vertraut wie beispielsweise Gymnasiasten.

Die Vertrautheit mit dem Computer beeinflusst die Einstellung zur Computer-Testung (Wildgrube, 1991); darüber hinaus beeinflusst sie die Motivation und den Grad der Leistung. Probanden, die keine Computer-Vorerfahrungen haben, sind benachteiligt, der Test entspricht dann nicht der erforderlichen „Fairness“.

Kubinger (1993) berichtet in dem Zusammenhang von einem interessanten Ergebnis einer Studie von Hergovich (1992): Wenn ein Lehrprogramm bezüglich der Handhabung des Systems (z.B. Benutzung der Maus zum Zeichnen am Bildschirm) der tatsächlichen Testung vorausgeht, ergaben sich keine signifikanten Mittelwertsunterschiede zwischen Probanden mit und ohne Computererfahrung (Hilke, 1993). Dieses Ergebnis spricht nach Kubinger (1993) für die „Erfahrungsunabhängigkeit“ der Computer-Testung. (Einen Überblick über angloamerikanische Studien bieten Klieme & Stumpf, 1990.)

Bezüglich der Akzeptanz von **Persönlichkeitstests** belegen ältere Studien - ebenso wie bei Leistungstests - eine erhöhte Ängstlichkeit unter Computer-Bedingungen. Neuere Studien bestätigen dies nicht (Klieme & Stumpf, 1990).

Im Vergleich zu Leistungstests erhalten Persönlichkeitstests unter Computer-Bedingungen eine höhere Akzeptanz. Einige Studien belegen sogar eine Bevorzugung der computerisierten Testversion (Klieme & Stumpf, 1990; Schwenkmezger & Hank, 1993).

Im allgemeinen wurde eine Reduzierung der Gesamttestdauer bei der Computer-Darbietungsform festgestellt. Schwenkmezger und Hank (1993) konnten aber in ihrer Studie diesen Befund nicht bestätigen.

Es wird von vielen Autoren angenommen, daß Probanden bei der Computer-Testvorgabe eine höhere Bereitschaft zeigen, ehrlich und offen zu antworten als bei der Papier-Bleistift-Testung. Nach der schon referierten Äquivalenz-Studie zum Gießen-Test von Bader, Hofmann und Kubinger (1993) konnte diese These nicht bestätigt werden. „Ob“ Probanden „es aber aufgeben, etwa bei einer Stellenbewerbung zu ihrem Vorteil zu antworten, ist mehr als fraglich“ (Kubinger, 1993, 134).

Zusammenfassung zum Exkurs

Die *Äquivalenz auf der Meßebe* (A) und die *Äquivalenz in der Akzeptanz der Probanden* (B) bezüglich Papier-Bleistift und Computer-Versionen wurde diskutiert. Die Präsentation erfolgte - wenn erforderlich - getrennt für die Leistungs- und Persönlichkeitstests. Dabei ist festzuhalten:

- Zu (A): Die *Psychometrische Äquivalenz* liegt bei Persönlichkeitstests höher als bei Leistungstests. Sie ist für jede Übertragung von Papier-Bleistift-Test auf den Computer zu prüfen.

In bezug auf die *Testgüte* profitieren von einer computerisierten Testvorgabe eher die Objektivität und die Ökonomie der Testung. Als sehr problematisch erweist sich das Kriterium der Normierung, während bei der Reliabilität, Validität und Nützlichkeit keine grundlegenden Unterschiede zwischen Papier-Bleistift- und Computer-Versionen vorliegen.

- Zu (B): Persönlichkeitstests erhalten im Vergleich zu Leistungstests unter Computer-Bedingungen eine höhere Akzeptanz. Als wichtige Einflußgrößen der Akzeptanz einer computerisierten Testvorgabe kann die Einstellung zu und die Vertrautheit mit dem Medium Computer angesehen werden.

Teil V

Integration: Multimodale Diagnostik und Intervention

Dieses Lehrbuch orientiert sich an der diagnostisch-interventiven Situation, diese Situation stellt den Psychologen vor unterschiedliche Forderungen:

- Er muß den **Gegenstandsbereich von Diagnostik und Intervention** überschauen *Teil I* umschreibt, was psychologische Diagnostik und Intervention besagen, und skizziert einige ihrer Modellvorstellungen.
- Der Psychologe sollte über bestimmte **Grundkenntnisse** verfügen. *Teil II* nennt Kenntnisse, die sich beziehen auf Testtheorie, Verhaltensbeobachtung und Gesprächsführung.
- Er muß mit **speziellen Einzelverfahren** vertraut sein, in deren Konstruktion die Grundkenntnisse eingegangen sind. Dazu zählt *Teil III* Leistungstests, Fragebögen und projektive Verfahren auf.
- Er sollte sich auskennen in bestimmten **Einzelfragen**. *Teil IV* führt Themen an wie Klassifikation und Selektion, Statistische und Klinische Urteilsbildung, Nutzenschätzung und edv-gestützte Diagnostik.
- Schließlich muß der Psychologe Techniken beherrschen, die es ihm erlauben, Informationen aus Grundkenntnissen und Einzelverfahren in besonderen Urteilsschritten **zu integrieren**: *Diese Vorgehensweisen soll Teil V darstellen.*

Integration bezeichnet hier erstens den Abruf von Informationen aus unterschiedlichen Wissensbereichen, zweitens ihre ‚simultane‘ Verwendung in der diagnostisch-interventiven Situation. Integration schließt damit jenes Vorgehen ein, das als ‚multimethodal‘ oder als ‚multimodal‘ bezeichnet wird, eine Prozedur, bei der ein Urteil aus Informationen gespeist wird, die mit unterschiedlichen Methoden gewonnen wurden.

Die integrativen Schritte seien in fünf Abschnitten vorgestellt:

- Wir sprechen vom *Untersuchungsverlauf* (Kap. 19).
- Wir skizzieren vier *Beispiele integrativer Diagnostik / Intervention*:
 - ⇒ Beispiel I: Antrag auf Therapieverlängerung (Kap. 20),
 - ⇒ Beispiel II: Psychologische Begutachtung (Kap. 21),
 - ⇒ Beispiel III: Beurteilung von Stellenbewerbern (Kap. 22),
 - ⇒ Beispiel IV: Assessment-Center (Kap. 23).

19. Kapitel

Zum Verlauf integrativer Diagnostik

Die diagnostische Untersuchung fordert dem Psychologen unterschiedliche Leistungen ab. Vereinfacht lassen sich drei Prozesse voneinander abheben (Leichner, 1979, 12):

- Am Anfang steht eine diagnostische Frage, die ein Proband formuliert.
- Am Ende soll der Psychologe eine Diagnose bieten: eine Aussage, welche die ‚diagnostische Frage‘ in umfassendere Verhaltenszusammenhänge einordnet und in diesem Sinne das Problem klärt.
- Aufgrund der Diagnose wird der Psychologe gegebenenfalls eine Prognose ableiten und interventive Maßnahmen vorschlagen.

An dieser vereinfachten Struktur seien einzelne Aspekte herausgehoben:

- Proband und Diagnostiker müssen *sich verständigen* (19.1).
- Einer Verständigung sind
allgemeine und spezielle Rahmenbedingungen gesetzt (19.2).
- Die ‚diagnostische Frage‘ ist in einen ‚psychologischen Kontext‘
zu übersetzen (19.3).
- Zur Problemlösung sind *Verfahren auszuwählen* (19.4).
- *Die Ergebnisse muß der Diagnostiker in einer Antwort integrieren* (19.5).
- Am Ende muß er dem Probanden seine *Antwort vermitteln* (19.6).
- Er wird, wenn nötig, *interventive Maßnahmen* vorschlagen (19.7).
- Er sollte den ‚Erfolg‘ seiner Vorschläge *evaluieren* (19.8).

Es folgen eine Zusammenfassung (19.9) und eine Reihe von Kontrollfragen (19.10).

19.1 Verständigungsaufgabe - Struktur des diagnostischen Urteils

Den diagnostischen Prozeß eröffnet eine Begegnung. Der Proband mit seiner Frage wendet sich an einen Diagnostiker mit seinem Fachwissen. Die beiden Personen müssen sich verständigen. An dieser Aufgabe läßt sich die formale Struktur des diagnostischen Urteils verdeutlichen.

Der Diagnostiker muß verstehen, was der Proband ihm mitteilen will. Für das Problem, das ihm mitgeteilt wird, soll er eine Lösung finden, gegebenenfalls aus seiner Antwort eine interventive Maßnahme ableiten.

Den Schritt des Verstehens nennen wir Interpretation - es ist eine rezeptive Leistung. Den Schritt der Lösungssuche nennen wir Hypothesentestung - es ist eine eher kreative Leistung.

Interpretation besagt, daß der Diagnostiker sich darüber klarzuwerden versucht, was der Proband ihm mitteilen will, und daß er sich bemüht, die Mitteilungen aus seinem Fachwissen zu ordnen und zu klären.

Die Interpretation schließt subjektive und objektive Elemente ein. „Die Subjektivität der Interpretation besteht darin, daß wir . . . eine Lebenssituation verstehen - über das hinaus, was sich der physikalischen Wahrnehmung erschließt. Die Objektivität der Interpretation jedoch besteht darin, daß dieses Verstehen keinen beliebigen Inhalt haben kann, sondern am gegebenen Gegenstand auch von anderen Subjekten nachvollzogen werden muß und kann“ (Seiffert, 1971 b, 135).

Interpretation schließt ihrerseits Hypothesen ein: Was ist eine **Hypothese**?

„Hypothese heißt wörtlich ‚Unterstellung‘, also soviel wie Annahme, Vermutung. Eine Hypothese ist eine Erklärung, mit der wir vorläufig arbeiten - solange, bis sie entweder erhärtet oder widerlegt ist“ (Seiffert, 1971 a, 139). Eine bestätigte Hypothese nennen wir ein Gesetz, allerdings nur dann, „wenn die Hypothese ein allgemeiner Satz ist und nicht nur eine Aussage über individuelle (Einzel-)Sachverhalte“ (Seiffert, 1971 a, 144).

Wenden wir diese Festlegungen (Definitionen) auf den (vereinfachten) Verlauf einer diagnostischen Untersuchung an.

Der Diagnostiker versucht, zwischen den Sachverhalten, die ihm der Proband mitteilt, und ihren ‚Ursachen‘ oder ihren ‚Bedingungen‘ Zusammenhänge zu erkennen. Die Aufgabe, größere Zusammenhänge darzustellen, fällt den Hypothesen zu: Hypothesen sollen begründete Erklärungen liefern für die vom Probanden mitgeteilten Einzelsachverhalte. Somit ergibt sich folgender **Erklärungszusammenhang**:

Dem Diagnostiker liegen vor:

- *einzelne Sachverhalte*, die ihm der Proband mitgeteilt hat (etwa: Person A hat starke Prüfungsängste);
- *allgemeine Aussagen* über Zusammenhänge: Hypothesen oder Gesetze (etwa: Wenn eine Person starke Leistungsängste hat, sinkt ihre kognitive Leistung ab);
- *eine Zuordnung*, die den konkreten Sachverhalt einordnet in allgemeine Zusammenhänge und ihn so ‚erklärt‘ (etwa: Weil Person A starke Prüfungsängste hat, bringt sie nur geringe kognitive Leistungen zustande).

Die Diagnose ist eine Interpretation einzelner Sachverhalte aus der Kenntnis umfassenderer Aussagen. Dabei kommt das **Grundwissen des Diagnostikers** ins Spiel. Denn der Diagnostiker interpretiert die Mitteilungen des Probanden

und formuliert über sie Hypothesen aufgrund des Sachwissens, das er sich erworben hat, etwa in der Allgemeinen, in der Entwicklungs-, in der Persönlichkeitspsychologie. Hinzu kommen die Erfahrungen, die er im Beruf gesammelt hat.

Das Insgesamt dieser Kenntnisse bildet ein Hintergrundwissen, das sich ständig verändert, erstens weil der Diagnostiker seine Kenntnisse individuell erweitert und zweitens weil das Wissen im Fach Psychologie zunimmt und der Diagnostiker an dieser Zunahme teilnehmen kann.

Prüfung von Hypothesen: Zu demselben Sachverhalt lassen sich meist unterschiedliche Zusammenhänge annehmen, also unterschiedliche Hypothesen erstellen. (Für die Prüfungsängste der Person A lassen sich vielerlei ‚Ursachen‘ oder ‚Bedingungen‘ annehmen.)

Die Aufgabe des Diagnostikers besteht darin, die Hypothesen nacheinander abzutesten. Der Prüfeffekt ist am günstigsten, wenn sich die Hypothesen als disjunkte Annahmen formulieren lassen.

Was aber wenn ein Untersucher es unterläßt, Hypothesen zu erstellen? Der Diagnostiker kann es nicht unterlassen, Hypothesen zu erstellen. Er kann es unterlassen, Hypothesen explizit zu formulieren, aber er kann nicht darauf verzichten, mit Hypothesen zu arbeiten. Auf Hypothesen zu verzichten hieße, auf Handeln zu verzichten.

Denn jede menschliche Handlung enthält Hypothesen - nicht nur das wissenschaftliche Handeln. Jede menschliche Handlung antizipiert eigenes und fremdes Verhalten, das in Zukunft eintreten könnte. Hypothesenbildung und Hypothesentestung kommen im Alltag ebenso vor wie in der Wissenschaft. Doch in der Wissenschaft werden Hypothesen gezielter gefaßt, klarer formalisiert.

Im diagnostischen Prozeß beruht die Formulierung von Hypothesen mit auf der Rezeption von Angaben, die der Proband macht, also auf Interpretationen. Darum verstehen wir den diagnostischen Untersuchungsverlauf als einen Vorgang, der sich zwischen Interpretation und Hypothesentestung bewegt.

19.2 Rahmenbedingungen der Psychologischen Situation

Das Treffen zwischen Proband und Diagnostiker zielt zwar auf Informationsaustausch, gestaltet sich aber auch als soziale Interaktion eigener Art (Burgoon & Ruffner, 1978; Dahmer & Dahmer, 1982; Graumann, 1972; Hartmann, 1984; Haubl, 1984; Jäger, R. S., 1986; Kisker, 1969; Scherer & Walbott, 1984; Schmidt, L.R., 1982, 1995).

Immer ist der Diagnostiker in einem finalen, sozialen, in einem ethisch-juristischen Kontext tätig.

Ethische und juristische Ansprüche sind der diagnostischen Arbeit immanent, sofern sie zu tun hat mit der Eigenständigkeit des Probanden und seiner Beziehung zu anderen Personen, also mit Selbst- und Fremdbestimmung, mit Freiheit und Personenwürde.

Davon war kurz die Rede in Kapitel 1, ausführlicher in Kapitel 12 (S. 8 und S. 337). Darum behandelt dieses Kapitel nur die psychologischen Rahmenbedingungen.

- Wir erwähnen zwei Arten:
- Allgemeine psychologische Determinanten

(19.2.1),

- Spezielle psychologische Determinanten

(19.2.2).

19.2.1 Allgemeine psychologische Determinanten

Allgemein sollen Determinanten heißen, welche jede soziale Interaktion mitbestimmen, auch die diagnostische Situation. Kasten 19-1 nennt Beispiele.

Mit einer Nennung ‚allgemeiner Determinanten‘ ist die diagnostische Situation noch nicht hinreichend beschrieben.

Kasten 19-1:
Allgemeine Determinanten einer diagnostischen Situation

<div><div>A) Randbedingungen:</div><div><div>- Raum-Dimensionen wirken sich auf die Begegnung aus, z. B.:</div><div><div>- Raumgröße, Raumgestalt;</div><div>- Helligkeit: Tageslicht, künstliches Licht, Lichtfarbe;</div><div>- Raumtemperatur;</div><div>- Art der Möblierung;</div><div>- Sitzverteilung;</div><div>- Störgeräusche.</div></div><div>- Zeit-Dimensionen beeinflussen die ‚Aussprache‘ mit, z.B.:</div><div><div>- der Termin der Begegnung (vom Probanden gewählt, vom Untersucher bestimmt?);</div><div>- die Dauer der Untersuchung (Wieviel Zeit steht zur Verfügung?);</div><div>- die Tageszeit, die Jahreszeit, der Biorhythmus (mit seinen Arbeits- und Ermüdungsphasen).</div></div></div></div>
<div><div>B) Gegebenheiten auf seiten des Probanden:</div><div><div>- Objektive Gegebenheiten des Probanden bestimmen den Verlauf mit, etwa:</div><div><div>- Geschlecht, Bildungsstand, Kleidung;</div><div>- Körpersprache (Gestik, Mimik).</div></div><div>- Subjektive Gegebenheiten des Probanden kommen ins Spiel, so:</div><div><div>- Ichbeteiligung, Begegnungsängste, Anfälligkeit für Streß;</div><div>- Freiwilligkeit der Untersuchung, Instruktionsverständnis, Perzeption des Untersuchers;</div><div>- Erwartungen an den Diagnostiker.</div></div></div></div>

C) Gegebenheiten auf seiten des Untersuchers

Auch beim Untersucher lassen sich objektive und subjektive Einflußgrößen ausmachen:

- Objektive Gegebenheiten wie
 - Geschlecht, Bildungsstand, Kleidung;
 - Körpersprache (Gestik, Mimik).
- Subjektive Gegebenheiten wie
 - Empathie, Engagement („sich selbst einbringen“);
 - Echtheit, Kompetenz, Diagnostik- und Therapieausbildung
(z. B. naturwissenschaftliche oder geisteswissenschaftliche Orientierung).

D) Interaktionen zwischen Proband und Untersucher:

- Gemäß *lerntheoretischen* Annahmen laufen Interaktionsprozesse ab wie wechselseitige Verstärkung (Lob und Tadel), Löschen/Hemmen.
- Gemäß *psychodynamischer* Interpretation ist mit Interaktionseffekten zu rechnen wie Übertragung und Gegenübertragung, Rollenspiel zwischen Diagnostiker und Proband.
- Der *Proband kann die Interaktion steuern*, beispielsweise durch Verweigerung der Mitarbeit, durch Lügen, Übertreiben, Untertreiben usw. (vgl. Hartmann, 1973, 58).
- Der *Untersucher kann die Interaktion steuern*, beispielsweise durch gezielte Lenkung des Gesprächs, durch Nachfragen bei Informationslücken, Aufklärung von Widersprüchen usw.

19.2.2 Spezielle psychologische Determinanten

Speziell heißen Determinanten, welche speziell die diagnostische Situation kennzeichnen (Spitznagel, 1982 a, 250-258).

Genannt seien:

- einseitige Offenheit,
- Bewußtsein der Beobachtung,
- vertrauliche Behandlung der Informationen,
- Vorgaben/Festlegungen.

Einseitige Offenheit: Vom Probanden wird erwartet, daß er über sich selbst und seine Probleme offen und aufrichtig Auskunft gibt. Umgekehrt wird dem Untersucher gestattet, nach sehr persönlichen, auch ‚intimen‘ Sachverhalten zu fragen.

In dieser „asymmetrischen Selbstenthüllung“ (Spitznagel, 1982 a, 251) wird der ethische Ansatz einer diagnostischen ‚Begegnung‘ erkennbar. Denn vom Probanden ist nur dann Offenheit zu erwarten, wenn ihr beim Diagnostiker Haltungen oder Einstellungen wie ‚Teilnahme‘ und ‚Hilfsbereitschaft‘ entsprechen.

Einschränkung: Das Prinzip ‚einseitiger Offenheit‘ gilt eher für Fälle, in denen der Proband den Psychologen aus eigenem Antrieb aufsucht. Wenn ein Proband ‚verpflichtet‘ wird, einen Psychologen aufzusuchen, läßt sich eine solche ‚Offenheit‘ nicht erwarten. Oder ist es Herrn X zu verübeln, daß er sich in einer psychologischen Untersuchung möglichst günstig darstellt, von

deren Ergebnissen es (mit) abhängt, ob er den Führerschein zurückbekommt, den er ,wegen Alkohols am Steuer‘ verloren hat?

Bewußtsein der Beobachtung: Die diagnostische Situation sieht vor, daß der eine beobachtet und der andere sich beobachten läßt. Nicht nur **sind** die Rollen asymmetrisch, die beiden Rollenträger sind sich der Asymmetrie auch **bewußt**. Beim Probanden kann dieses ‚reflexive‘ Moment Angst auslösen, beim Diagnostiker Machtgefühle wecken.

Vertrauliche Behandlung der Informationen: Zu dem Rollensystem Klient/Diagnostiker gehört die Erwartung und die Zusicherung von Verschwiegenheit. Nur unter dieser Voraussetzung wird der Proband ‚persönliche‘, auch intime Informationen preisgeben, nur unter dieser Voraussetzung darf der Diagnostiker sie annehmen.

Wem ist Verschwiegenheit auferlegt? Dem Untersucher selber, aber auch seinen Mitarbeitern und den Kollegen, mit denen er den Fall bespricht!

Vorgaben/Festlegungen: Unter vielen Aspekten ist der Diagnostiker festgelegt, so sehr, daß er Vorgaben und Einschränkungen oft nur registrieren, nicht eliminieren kann (Kaminski, 1970, 401-425). Kasten 19-2 führt Beispiele an.

Kasten 19-2:
Konkrete Vorgaben und Festlegungen für den Diagnostiker

Zeitliche Vorgaben:

- Terminkalender sowohl des Probanden wie auch des Untersuchers.

Institutionelle Vorgaben:

- Technisch-organisatorischer Rahmen (Räumlichkeiten, Dienstpläne),
- Repertoire an Arbeitsmitteln,
- Personelle Vorgaben: Zahl, Geschlecht, Kompetenz der Mitarbeiter.

Festlegungen durch den Auftraggeber:

- Begrenzung der Untersuchung nach Umfang und Zeit,
- Begrenzung der Kosten.

‚Prägung‘ des Diagnostiker durch seine Biographie, vor allem

- durch die psychologische ‚Schule‘, die er absolviert hat,
- durch die Denkmodelle, die ihm geläufig sind,
- durch die ‚Untersuchungsroutinen‘, auf die er sich verläßt.

Bei dem Kontakt mit dem Probanden sind dem Untersucher also Rahmenbedingungen vorgegeben. In diesem Rahmen muß er die Fragen des Probanden entgegennehmen und interpretieren.

19.3 Übersetzungsprobleme

Der Kontakt hat ein Ziel: Der Proband soll seine Probleme beschreiben, der Untersucher sie verstehen. Dabei seien zwei Sequenzen unterschieden:

- *Angemessene Fragestellung:*
Artikulierte der Proband seine Frage angemessen (19.3.1)?
- *Angemessene Übersetzung der Fragestellung:*
Überträgt der Untersucher die Probanden-Frage
angemessen in einen psychologischen Kontext (19.3.2)?

19.3.1 Angemessene Fragestellung durch den Probanden

Der Proband kommt, um dem Diagnostiker seine Probleme vorzutragen. Dabei kann es sich um manifeste oder um verschlüsselte Fragen handeln.

Als **manifest** soll eine Frage gelten, wenn der Proband sein Problem explizit benennen und beschreiben kann.

Beispiele:

1. *Eignungsuntersuchung: Bin ich für den Beruf des Schreiners geeignet?*
2. *Schullaufbahn: Soll ich zur Realschule oder zum Gymnasium gehen?*

Gemeint ist mit manifesten Fragen, daß klar ist, wonach gefragt wird. Nicht gemeint ist, daß die Klärung der Frage selber leicht sei.

Als **verschlüsselt** soll eine Frage gelten, wenn der Proband sein Problem gleichsam ‚versteckt‘. Er erscheint bei dem Psychologen unter einem Vorwand, den eigentlichen Grund nennt er nicht, kennt ihn oft auch nicht.

Beispiel: *Eltern suchen mit ihrem sechzehnjährigen Sohn (dem Probanden) und seinem fünfzehnjährigen Bruder einen Psychologen auf. Wie im obigen Beispiel geht es - scheinbar - um Berufsberatung. Aber im Verlaufe der Vorgespräche stellt sich heraus: Die ‚Berufssuche‘ war nur der **Anlaß**, nicht der **Grund** für den Gang zum Berater: Hinter der ‚Berufsfrage‘ verbirgt sich eine Fülle anderer Probleme des Jungen: Suche nach der sozialen, speziell seiner sexuellen Rolle; Unsicherheit in der Beziehung zu Mädchen; Leistungsschwierigkeiten in der Schule. - Hier lautet die manifeste Frage: Für welchen Beruf ist unser Sohn/bin ich geeignet? Aber die ‚eigentliche‘ Frage ist fundamentaler*

Verschlüsselte Fragen müssen in der Untersuchung überhaupt erst als solche erkannt und artikuliert werden. Ob die ‚entschlüsselte Frage‘ dann zum Gegenstand der Untersuchung gemacht wird, darüber müssen Diagnostiker und Proband in einer Aussprache neu befinden.

Die beiden Frageklassen (manifest, verschlüsselt) lassen sich nicht disjunkt trennen, sie liegen auf einem Kontinuum. Die Unterscheidung erinnert nur daran, daß eine vorgebrachte Frage nicht die ‚echte‘ Frage sein muß, somit die Erarbeitung der ‚eigentlichen Frage‘ zur Untersuchungsaufgabe gehören kann.

19.3.2 Angemessene Übersetzung durch den Diagnostiker

Meist trägt der Proband sein Anliegen in der Umgangssprache vor. Der Psychologe muß das Problem in die Fachsprache übertragen (Jäger, R. S., 1986, 65). Diese Übersetzung schließt ein, daß er seine Ergebniserwartungen, seine Hypothesen, artikuliert.

Beispiele:

1. *Eine Mutter fragt: Ist unser Sohn schulreif? Der Psychologe muß die globale Frage aufschlüsseln und beispielsweise (als Hypothese) prüfen, ob das Kind über eine kognitive, emotionale, motorische, soziale ‚Ausstattung‘ verfügt, die den Schulbeginn rechtfertigt.*
2. *Der Sechzehnjährige in dem Beispiel ‚verschlüsselte Frage‘ bittet (zunächst) um Hilfe bei der Berufswahl. Schon dies ist eine umfassende Frage, bei der sich die Suche nicht auf das engumschriebene Gebiet von Eignung und Neigung beschränken kann, sondern ein breiteres Feld abstecken muß. Welches Wissen ist abzurufen?*
 - *Berufswahltheorien,*
 - *Motivationstheorien,*
 - *Kenntnisse über Eignungsmessung,*
 - *Kenntnisse über Anforderungsprofile,*
 - *Erkenntnisse über Lebensplanung*
 - *usw.*
3. *Daß der Sechzehnjährige nicht allein kommt, sondern von Eltern und Bruder begleitet wird: dieses Verhalten wirft eigene Fragen auf*
 - *Beteiligen sich Eltern und Bruder nur uneigennützig an der Berufssuche von Sohn/Bruder? Oder geht es ihnen um ‚Kontrolle‘ der Berufswahl, um Rollen- und Machtverteilung in der Familie, um Wege einer Ablösung vom Familienverband?*
 - *Liegt hinter der Sorge der Eltern um den Berufsweg des Sohnes die umfassendere Sorge um seinen Platz im Leben? War dies der Grund, der sie zum Psychologen trieb?*

Indem der Diagnostiker die Ausgangsfrage zerlegt und psychologisch neu strukturiert, entwickelt er aus den interpretierten Aussagen des Probanden zugleich seine (Ergebniserwartungen, seine) Arbeitshypothesen. Dabei dürfte die Regel sein, daß Fragestellung und Untersuchungsplan erst im Fortgang der Untersuchung schärfere Umrisse gewinnen.

19.4 Verfahrensauswahl/Korrespondenzprobleme

Das Problem sei identifiziert und in einen psychologischen Kontext übersetzt. Damit ergibt sich auch die Aufgabe, das Untersuchungs-Instrumentarium auszusuchen. Es stellen sich zwei Fragen:

- *Eine Zuordnungsfrage:*
Welches Problem wird mit welchem Instrument untersucht? (19.4.1).
- *Eine Anordnungsfrage:*
In welcher zeitlichen Sequenz werden die Verfahren eingesetzt? (19.4.2).

19.4.1 Zuordnungsfrage:

Korrespondenz von Problem und Verfahren

Vor die Zuordnungsfrage stellt den Psychologen folgender Tatbestand: Konkrete diagnostische Fragestellung und konkretes diagnostisches Instrument sind einander nicht (gleichsam a priori) zugeordnet, ihre ‚Passung‘ ergibt sich nicht ‚von selber‘, sie muß ‚hergestellt‘ oder ‚festgestellt‘ werden.

Die meisten diagnostischen Instrumente sind für mehrere Problemklassen verwendbar. So läßt sich der ‚Intelligenz-Struktur-Test‘ (IST 70: Amthauer, 1973) ebenso bei einer Eignungsuntersuchung (Soll der Proband einen akademischen Beruf ergreifen?) wie bei einer forensischen Fragestellung einsetzen (Hat der Täter die Umstände seiner Tat erfaßt?).

Ob ein konkretes Verfahren dazu beiträgt, eine diagnostische Frage zu beantworten, muß der Untersucher selber feststellen. Anders gesagt: Da keine generellen Regeln vorliegen, welche die Korrespondenz zwischen konkreter diagnostischer Frage und ihrer ‚Abbildung‘ in einem konkreten Instrumentarium festlegen, ist der Psychologe auf seine Fachkenntnis, seine Erfahrung und seine Intuition angewiesen - und auf den Rat seiner Kollegen.

Beachten sollte er dabei auch systemische Aspekte, er sollte also bestrebt sein, die Struktur der Beziehungen zu erkennen, in die ein Proband gestellt, auch wohl ‚eingebunden‘ ist: seine Stellung im ‚System‘ der Familie, im System der Gleichaltrigen, des Arbeitsteams usw.

Um die Aufgabe zu veranschaulichen, sei das Beispiel der **‚Berufssuche des Sechzehnjährigen‘** weitergeführt:

- *Das Gesamtproblem kann der Untersucher in Teilprobleme zerlegen. Das Gesamtproblem ‚Berufssuche‘ läßt sich etwa aufgliedern in Teilbereiche wie: Subjektive Angaben (Selbstbild) zu Eignungen und Interessen, objektive Angaben zu Neigungen und Eignungen, Sichtung der Lerngeschichte, Antizipation ‚persönlicher‘ und beruflicher Zukunft, Facetten von Leistungsmotivation, Stellung im familiären Verband usw.*

Für die einzelnen Teilbereiche muß der Untersucher Hypothesen artikulieren und ihnen ‚passende‘ Instrumente zuordnen:

- ⇒ *Selbstbild von Interessen und Fähigkeiten beeinflusst Berufswahl und Berufssuche. Erfassung mit Exploration, mit ‚Berufs-Interessen-Test‘ (BIT: Irle, 1955), mit ‚Differentiallem Interessen-Test‘ (DIT: Todt, 1967) usw.*
- ⇒ *‚Objektive‘ Fähigkeitsmaße helfen, die ‚richtige‘ Berufssparte zu entdecken. Erfassung mit Leistungstests, z. B. ‚Berufs-Eignungs-Test‘ (BET: Schmale & Schmidtke, 1966) usw.*
- ⇒ *Vergangene Lerngeschichte beeinflusst gegenwärtige Interessen und Neigungen. Erfassung mit Exploration, mit biographischem Fragebogen, gegebenenfalls dem ‚Biographischen Inventar zur Diagnose von Verhaltenstörungen‘ (BIV: Jäger R.S. & Mitarb., 1976) usw.*
- ⇒ *Antizipation der Zukunft ‚verrät‘ fundamentale Richtungen von Interessen, Neigungen, Fähigkeiten. Erfassung mit Exploration, mit projektiven Verfahren usw.*
- ⇒ *Berufserfolg setzt Leistungsmotivation voraus: Für welche Bereiche läßt sich Leistungsmotivation erkennen? Erfassung mit dem ‚Thematischen Apperzeptions-Test‘ (TAT), mit dem ‚Leistungsmotivations-Gitter‘ (LM-Gitter: Schmalt, 1976), mit Interviews usw. Gegebenenfalls muß er aus der TAT-Serie Tafeln aussortieren, die ‚versprechen‘, Informationen über Leistungsmotivation abzuwerfen.*
- *Für Teilbereiche, die der Diagnostiker in einer Exploration berühren will, muß er Themenbereiche abgrenzen (vielleicht auch Einzelfragen vorbereiten).*
- *Er muß Leistungstests auswählen, welche für die anstehende Fragestellung relevante Merkmale erfassen: Intelligenz, Handgeschick usw.*
- *Er muß Persönlichkeitsinventare aussuchen, die für die anstehende Fragestellung relevante Persönlichkeitsmerkmale erfassen: Extra-/Introversion usw.*

Die Veranschaulichung dürfte verdeutlichen, daß die drei Teilaufgaben (Aufgliederung des Gesamtproblems, Artikulation der Hypothesen und Auswahl relevanter Verfahren) miteinander zusammenhängen, darum auch auseinander hervorgehen: in einem Prozeß, der differenzierter verläuft, als das Beispiel zeigen kann.

Sonderproblem: *Spezielle diagnostische Aufgaben können den Diagnostiker in eine Lage versetzen, in der er feststellen muß, daß er kein geeignetes Instrumentarium vorfindet. Dann hat er zu entscheiden, ob ihm ‚sekundäre‘ Informationsquellen wie Zeugnisse, Vorgesetzten- oder Kollegenurteile usw. genügen. In Sonderfällen muß er aber auch befinden, ob er neue Verfahren entwickeln soll und, wenn ja, ob er es kann (vgl. die Entwicklung des Tests von Binet und Simon, des sogenannten ‚Binetariums‘ oder des ‚Tests für medizinische Studiengänge‘: Kap. 1, S. 6).*

Gelegentlich wird der Diagnostiker sich fragen müssen, ob er eine ‚diagnostische Frage‘, die ihm gestellt wurde, mit psychologischen Mitteln überhaupt lösen kann.

19.4.2 Anordnungsfrage: Korrespondenz von Problem und Sequenz der Verfahrensvorgabe

Wenn geklärt ist, wie diagnostische Frage und Instrumente einander zuordnenbar sind, dann ist die zeitliche Reihenfolge zu bestimmen, in der die Verfahren eingesetzt werden. Da eine ‚optimale Reihenfolge‘ von Verfahren nicht bekannt ist, lassen sich nur Hilfen im Sinne von Faustregeln formulieren (Bierkens, 1968, 55; Durchholz, 1981, 276).

Der Zeitplan sollte so konzipiert werden, daß er den Probanden für eine Mitarbeit gewinnt und Rückmeldungen darüber erlaubt, wie die jeweiligen Ergebnisse zu den Hypothesen ‚passen‘. Damit sind zwei Perspektiven genannt:

- Akzeptanz des Verlaufs durch den Probanden und
- Ergebnisbewertung durch den Untersucher.

Akzeptanz durch den Probanden

Um den Probanden für die Mitarbeit zu gewinnen, empfehlen sich Strukturierungen wie die folgenden:

- An den Beginn sollten Verfahren rücken, die der Proband auf sein Problem beziehen kann (Bedeutsamkeit der Augenscheinvalidität).

Beispiele:

1. Bei der Fragestellung ‚Berufswahl‘ sollte ein Gespräch die Untersuchung eröffnen, das die Berufs- und Interessensfragen abzuklären sucht.
 2. Bei Kindern können zeichnerische Verfahren Eisbrecher-Funktionen übernehmen: ‚Mann-Zeichen-Test‘ (MZT: Ziler, 1975), ‚Baum-Test‘ (Koch, 1972), ‚Familie in Tieren‘ (Brem-Gräser, 1975).
- Die Abfolge der Verfahren sollte einen Rhythmus vorsehen: etwa einen Wechsel zwischen ‚geschlossenen‘ Verfahren, die dem Probanden wenig Spielraum gewähren (z. B. Leistungstests), und ‚offeneren‘ Verfahren, die ihm mehr Spontanität erlauben (z. B. Gespräche, Projektive Verfahren).
 - Die Befindlichkeit des Probanden (Ermüdung, Widerstand gegen ein Verfahren usw.) sollte jeweils den Verlauf mitbestimmen, Zeit- und Untersuchungsplan gegebenenfalls geändert werden.
 - Bei längerer Dauer sollte der Untersuchungsplan Erholungspausen für den Probanden vorsehen.

Ergebnisbewertung durch den Untersucher

Die zeitliche Reihenfolge der Verfahren orientiert sich nicht nur am Probanden, sie muß auch den Zielen des Diagnostikers dienen, vor allem seiner Aufgabe, Einzel- und Gesamtergebnisse in ihrem Wert für seine Hypothesen einzuschätzen.

- Wenn die Fragestellung selber noch zu klaren ist (immer bei ‚verschlüsselten‘, oft auch bei ‚manifesten‘ Fragen), dürfte es sich empfehlen, die Instrumente so anzuordnen, daß sie trichterförmig das Problem eingrenzen: zuerst Breitbandverfahren (bezogen auf die konkrete Frage), dann speziellere Verfahren.
- Es ist nützlich, den Verlauf so zu planen, daß der Diagnostiker schon während der Untersuchung Ergebnisse gewinnt und evaluiert. Zwischenergebnisse und Zwischenbewertungen ermöglichen es ihm, den Ablauf zu ändern, ihn zu verkürzen oder zu erweitern.
- Wie für den Probanden soll der Untersucher deshalb auch für sich selber Pausen vorsehen, mit dem Ziel, Informationen zu evaluieren (und, wenn möglich, mit Kollegen zu besprechen).
- Es kann sinnvoll sein, Zwischenergebnisse nicht nur mit Kollegen, sondern auch mit dem Probanden zu besprechen. Dies wurde bedeuten, die ‚Reaktivität‘ psychologischer Meßprozeduren diagnostisch auszunützen und eine Art ‚interaktiver Diagnostik‘ zu entwickeln.

Das Zusammenspiel von (inhaltlicher) Zuordnung und (zeitlicher) Anordnung könnte sich in einem Zeitplan abbilden, der die Untersuchungssequenz festlegt.

19.5 Integration der Ergebnisse

Die Untersuchung zielt darauf, aus den Ergebnissen jene Aussagen zu gewinnen, die eine Antwort auf die diagnostische Frage erlauben.

Inhaltlich vollzieht sich diese Ableitung während des gesamten diagnostischen Prozesses. Abgeschlossen wird sie **nach** der Untersuchung. Dieser Abschluß versteht sich in dem Sinne, daß die ‚benötigten‘ Informationen zusammengekommen sind (zusammengekommen sein sollten) und darum die Suche beendet wird.

Formal strukturiert sich dieser Integrationsprozeß immer wieder unter zwei Perspektiven: Zu *Teilfragen* müssen Aussagen integriert, diese dann um die *Gesamtfrage* zentriert werden.

Beispiel: *Bei der Berufssuche des Sechzehnjährigen muß der Psychologe sich darüber klar werden, wann Teilbereiche wie ‚Interessen‘ und ‚Fähigkeiten/Fertigkeiten‘ adäquat erfaßt sind, wann er also erkennen kann, ob ‚Interessen‘ und ‚Fähigkeiten‘ des Probanden einander entsprechen und sich ergänzen. Erst aus solchen ‚Vorklärurigen‘ entwickelt sich eine ‚Gesamtentscheidung‘, die etwa umreißt, welche Berufe der Proband in Erwägung ziehen und welche er außer Betracht lassen sollte.*

In diesem Beispiel müßte der Diagnostiker auch entscheiden, ob die Berufswahl selber als ‚Teilfrage‘ eines umfassenderen Such- und Beratungsprozesses zu behandeln ist. Dieser ‚neuen‘ (umfassenderen) Untersuchung müßte der Proband überhaupt erst zustimmen.

Der Diagnostiker muß die Teil- und Gesamtergebnisse evaluieren und ‚entscheiden‘, ob seine Hypothesen bestätigt worden sind.

Im Dienste dieser Entscheidungsfindung stehen vor allem zwei Klassen von Urteilsmodellen: statistische und klinische Entscheidungsstrategien. An dieser Stelle sei nur der Grundgedanke skizziert (ausführlicher werden sie besprochen in Kapitel 14, S. 351).

Wenn Daten quantifiziert vorliegen (etwa Test- oder Fragebogen-Scores usw.) und ihre Kombination auf einem ausformulierten Algorithmus beruht, dann spricht man von *statistischer Urteilsbildung*.

Wenn quantitative und qualitative Daten vorliegen (etwa Test-Scores, Zeugnisse, Verhaltensbeobachtungen usw.) und ihre Kombination auf dem Fachwissen, der Erfahrung, der Intuition des Diagnostikers beruht (ohne daß die Regeln des Urteilsganges mit allen Elementen explizit genannt werden), dann spricht man von *klinischer Urteilsbildung*.

Der Integrationsprozeß der Entscheidung muß nicht bei einer einzigen Antwort enden. Die Antwort kann ein mehrschichtiges Aussagengefüge erfordern. Zum Beispiel kann eine gutachterliche Antwort sich aufgliedern in Aussagen über Zahl und Struktur von Störungen, in Aussagen über deren (‚Ursachen‘ oder) Bedingungen und in Vorschläge für eine Therapie.

Am Ende sollte die Antwort auch formal der Frage entsprechen: Eine dreigliedrige Frage sollte eine (mindestens) dreigliedrige Antwort erhalten. Es geht dabei nicht um eine formale Entsprechung, sondern darum, daß die Antwort alle Aspekte der Frage berücksichtigt.

Beispiel:

Heiner R. (18;4 Jahre) kam auf Empfehlung des Schulleiters und auf Wunsch seiner Eltern zu einer psychologischen Untersuchung. Vor zwei Jahren hatte er die zehnte Gymnasialklasse wiederholt, er besucht jetzt die elfte Klasse, die Versetzung ist gefährdet. Wird er nicht versetzt, muß er die Schule verlassen,

Es werden folgende Fragen gestellt:

- 1. Ist Heiner in der Lage, das Gymnasium mit dem Abitur abzuschließen?*
- 2. Wenn ja, ist es sinnvoll, daß er an einer anderen Schule in die elfte Klasse wechselt?*
- 3. Für den Fall, daß er das Schulziel nicht erreicht: läßt sich schon jetzt klären, welchem Beruf er sich zuwenden könnte?*

Bei einer solchen dreigliedrigen Anfrage sollte die Antwort ebenfalls mindestens dreigliedrig ausfallen,

Wie bei jedem einzelnen Untersuchungsschritt, muß der Diagnostiker auch bei dem letzten Schritt der Integration **mit** all jenen **Fehlern rechnen**, die unter dem Titel der impliziten Persönlichkeitstheorie zusammengefaßt wurden, also mit Hofeffekt, Mildeeffekt usw. (vgl. S. 199). Zu diesen 'üblichen' Verzerrungstendenzen kann aber als besonderer Fehler hinzukommen, daß er die Informationen nicht ausschöpft, die er gesammelt hat (vgl. Barendregt, 1961; Leichner, 1975, 1976; Pelzmann, 1972).

19.6 Vermittlung der Antwort an den Probanden

Die Antwort(en), die der Untersucher gefunden hat, muß er dem Probanden (oder Auftraggeber) vermitteln.

Wenn gesagt wird, die Antwort richte sich an den Probanden (oder Auftraggeber), so ist das als verkürzte Darstellung gemeint. Adressat der Antwort muß nicht der untersuchte Proband, sondern kann auch eine andere Person sein, beispielsweise die Eltern, ein Richterergremium, eine Behörde.

Zu vermitteln hat der Diagnostiker die ganze Vielfalt der Gehalte - die kognitiven, emotionalen, behavioralen Facetten von diagnostischer Frage und Antwort. Aus der psychologischen Fachsprache muß er seine Aussagen rückübersetzen in die Sprache des Adressaten: eine Interpretationsaufgabe, die im Blick auf den Empfänger zu lösen ist. Demgemäß sollte die Antwort für einen Apotheker anders formuliert werden als für einen Schreinermeister.

19.7 Intervention

Die Antwort, die der Diagnostiker dem Adressaten gibt, schließt in der einen oder anderen Weise einen Interventionsvorschlag ein. Dieser Vorschlag muß spezifisch ausfallen: spezifisch *zum einen* für die Teildisziplin, in der die diagnostische Aufgabe zu lösen ist, spezifisch *zum anderen* für die Person oder Personengruppe, der die diagnostische Arbeit gilt.

Solche Interventionsanregungen „müssen schlüssig an die diagnostischen Befunde anknüpfen und dem aktuellen Stand der Forschungen entsprechen“ (dpv: Richtlinien, 1994 b, 8).

Drei Beispiele:

1. In der Pädagogischen Psychologie habe eine Untersuchung eine mehrgliedrige Störung festgestellt haben (etwa: Angst, Überforderung durch die Eltern, Leistungsruckstände). Eine Intervention müßte dann ebenfalls mehrgliedrig angelegt werden. Erstens sind Techniken einzusetzen, welche die Angst reduzieren. Zweitens ist auf die Eltern einzuwirken (Elternges-

sprache) mit dem Ziel, eine stützend-akzeptierende Haltung zu wecken. Drittens sollte bedacht werden, wie sich Lernschwächen auf lange Sicht ausgleichen lassen (Lorenz, 1987, 93-103).

2. *In der **Forensischen Psychologie**, im Vollzugsdienst, habe ein Psychologe die Prognose zu treffen, wie wahrscheinlich es ist, daß ein bestimmter Proband wieder rückfällig wird und welche Maßnahmen geeignet sein könnten, einen Rückfall zu verhindern. Ein Vorschlag - etwa Urlaub zu gewähren - muß aus der Individualprognose hervorgehen, die der Diagnostiker dem Anstaltsleiter unterbreitet (Liebel & Uslar 1975, 141-145).*
3. *In der **Arbeits- und Organisationspsychologie** habe die Diagnose eines Produktionsbetriebes ergeben, daß der Arbeitsablauf ineffektiv organisiert ist. Diesem Befund müssen die Interventionsvorschläge entsprechen: etwa Änderung von Arbeitsabläufen, Änderung von Anlertechniken, Umsetzung bestimmter Mitarbeiter Neuverteilung von Verantwortlichkeiten (Becker & Langosch, 1990, 88).*

19.8 Erfolgskontrolle: Evaluation und Supervision

Wenn möglich, sollte der Diagnostiker - nach einem angemessenen Zeitraum - feststellen, ob seine Diagnose, seine Ratschläge, seine Interventionsvorschläge ‚richtig‘ waren und sollte die Ergebnisse mit Kollegen besprechen, er sollte sich einer permanenten kollegialen Konsultation und vorher einer formellen Supervision unterziehen.

Psychologische *Diagnostik als Wissenschaft* darf auf solche Evaluation nicht verzichten. *Im Einzelfall* jedoch kann sich eine Erfolgskontrolle als unmöglich erweisen.

Beispiel:

1. *Ein Sorgerechtsfall, Ein Paar mit zwei minderjährigen Kindern lasse sich scheiden. Vor dem Familiengericht beanspruche jeder Elternteil das Sorgerecht für sich allein. Ob ein Richter jenem Elternteil das Sorgerecht zuspricht, den der Psychologe als Gutachter vorgeschlagen hat, läßt sich leicht feststellen. Kann dies schon heißen, den ‚diagnostischen Effekt‘ zu identifizieren? Gehört dazu nicht auch, zu verfolgen, was nach der richterlichen Entscheidung aus Kind und Eltern wird? Genau eine solche Bemühung kann sich aber dann als ‚erfolglos‘ erweisen, wenn die betroffenen Personen keinen Kontakt zum ‚Gutachter‘ mehr wünschen.*
2. *Eine Fahrtauglichkeitsuntersuchung. Ein Fahrer habe den Führerschein wegen ‚Alkohols am Steuer‘ verloren. Zur Entscheidung stehe, ob er nach einer Wartezeit die Fahrerlaubnis zurückerhält. Ein Gutachten werde angefordert. Rät nun der Gutachter zur Fortsetzung des Entzuges und folgt die ‚Verwaltung‘ diesem Vorschlag: woran bemißt sich dann der ‚diagnostische Erfolg‘? Daran, ob das zuständige Amt dem Entscheidungsvor-*

schlag des Diagnostikers folgt? Oder daran, wie sich der Proband im Verlaufe seiner weiteren Fahrerkarriere verhält?

19.9 Zusammenfassung zu Kapitel 19

Kapitel 19 sei zusammengefaßt in einem Schema, das noch einmal die Vieldimensionalität eines Untersuchungsverlaufes veranschaulichen soll (Kasten 19-3).

Kasten 19-3:
Wiederholung - Vereinfachte Darstellung eines Untersuchungsverlaufs

Fragestellung	Proband und Diagnostiker im ersten Gespräch.
Verständigung	Interaktion zwischen Probanden und Diagnostiker unter besonderen Randbedingungen (z.B. Einseitige Offenheit)/Ethische Implikationen (z. B. Vertraulichkeit).
Übersetzungsprobleme	Interpretation: Übersetzung der Aussagen des Probanden in die psychologische Fachsprache. Hypothesentestung: Erklärung des diagnostischen Problems durch Zuordnung der Einzelsachverhalte in umfassendere Zusammenhänge.
Korrespondenzaufgabe	Sachplan: Zuordnung des Problems zu relevanten Verfahren. Zeitplan: Anordnung der Verfahren in einer sinnvollen Sequenz.
Untersuchung	Vorlage der Verfahren.
Auswertung	Quantitative und qualitative Aufschlüsselung der Verfahren.
Integration	Aufgliederung der Antworten nach Teilthemen/Zusammenfassung der Teilantworten zu einem Aussagengefüge/Erarbeitung einer Stellungnahme.
Rückübersetzung/Beratung	Vermittlung der Ergebnisse an den Probanden./Darum Rückübersetzung der Befunde aus der psychologischen Fachsprache in die Sprache des Probanden.
Intervention	Die Ergebnisse führen zu Schritten, die das Wohlbefinden des Probanden erhöhen oder wiederherstellen sollen.
Evaluation/Supervision	Kontrolle des untersuchten Falles/der Ergebnisse durch den Diagnostiker/durch Fachkollegen. Besprechung/Diskussion von Fällen und Ergebnissen mit Fachkollegen.

19.10 Kontrollfragen zu Kapitel 19

- Determinanten der diagnostischen Situation.
- Übersetzung der Frage des Probanden in einen psychologischen Kontext.
- Zuordnungsfragen.
- Anordnungsprobleme.
- Integration von Einzelbefunden.
- Vermittlung der Antwort an den Fragesteller.

20. Kapitel

Beispiel I für Integrative Diagnostik ***Antrag auf Verlängerung einer Psychotherapie***

Thomas Fuchs

Kapitel 20 bringt ein erstes Beispiel integrativer Diagnostik und Intervention. Das Beispiel wurde ausgewählt, weil es an einem Einzelfall verdeutlicht, wie unmittelbar die diagnostischen Schritte in interventive Maßnahmen übergehen können: Diagnostisches Instrument ist das Interview - im Dienst einer Psychotherapie.

Das Kapitel wird in drei Abschnitte unterteilt:

- Vorbemerkung (20.1),
- Einführung (20.2),
- Text des Verlängerungsberichts (20.3).

20.1 Vorbemerkung

Der vorliegende Antrag auf Therapieverlängerung bezieht sich auf einen „echten Fall“. Die personenbezogenen Daten und Angaben wurden so verändert, daß es unmöglich ist, die Klientin zu identifizieren.

Der Autor ist ausgebildet in „Gestalttheoretischer Psychotherapie“. Die Darstellung der therapeutischen Maßnahmen orientiert sich am Konzept einer „methodenübergreifenden Psychotherapie“. Das Schema eines solchen Antrages und der Wortlaut der jeweiligen Überschriften entspricht einem Vordruck, den der *Berufsverband Deutscher Psychologen e.V. (BDP), Vertragsabteilung Psychotherapie*, herausgibt.

Gemäß dem BDP-Vordruck erhält der Antrag die Bezeichnung „Verlängerungsbericht“.

20.2 Einführung

Krankenkassen übernehmen die Therapiekosten nur für einen begrenzten Zeitraum: in der Regel für eine bestimmte Anzahl an Behandlungen (Sitzungen). Sind nach Ablauf der bewilligten Zahl von Behandlungen Psychotherapeut und Klient der Meinung, daß die Psychotherapie weitergehen soll, so beantragen sie eine Verlängerung.

Der folgende Text ist der Verlängerungsbericht eines Psychotherapeuten an einen Gutachter, der den Verlängerungsbericht unter drei Gesichtspunkten beurteilt:

- Notwendigkeit (Liegt eine psychische Störung mit Krankheitswert gemäß Sozialgesetzbuch [SGB] vor?)
- Zweckmäßigkeit (Ist die Behandlung dem Störungsbild angemessen und läßt sie einen Behandlungserfolg erwarten?)
- Wirtschaftlichkeit (Steht der Behandlungsumfang in angemessener Relation zum Therapieergebnis bzw. Therapieziel?).

Wenn der Gutachter mit dem Bericht einverstanden ist, empfiehlt er der Krankenkasse, die Kosten für eine Fortführung der Behandlung zu übernehmen.

Aus Gründen des Datenschutzes erfährt der Gutachter den Namen des Klienten nicht. Zur korrekten Zuordnung dient der Patientencode, der sich zusammensetzt aus zwei persönlichen Daten des Klienten: (1) dem ersten Buchstaben seines Familiennamens, (2) seinem Geburtsdatum.

20.3 Text des Verlängerungsberichts

Verlängerungsbericht vom 19.5.1993

des Therapeuten

Klinischer Psychologe/Psychotherapeut BDP

X Y Z

an die Gutachter der

Vertragsabteilung Psychotherapie (VAP) im BDP

Patientencode: *F010255*

1. Zur Person

Geschlecht: *weiblich*

Alter: *37 Jahre*

Familienstand: *in Partnerschaft lebend*

Kinderzahl: *1*

Alter des Kindes: *7 Jahre*

Schulabschluß: *Abitur*
 Erlernter Beruf: *Kaufmännische Angestellte*
 Jetzige Tätigkeit: *Hausfrau und Mutter Kaufmännische Angestellte*
(1/2 Stelle)

2. Daten zur bisherigen und geplanten Behandlung

2.1 Bisherige Behandlung

Behandlungsbeginn: 18.09.1992
 Erstantrag: 26.10.1992
 Anzahl der durchgeführten Behandlungen: 25 *Behandlungen*
 Anzahl der bewilligten Sitzungen: 25 *Behandlungen*
 Frequenz der Behandlungen: 1 *Behandlung wöchentlich*

2.2 Geplante Behandlung

Anzahl der neu beantragten Behandlungen: 50 *Behandlungen*
 Beabsichtigte Frequenz der Behandlungen: 1 *Behandlung wöchentlich*

3. Basisdaten zum Zeitpunkt des Behandlungsbeginns

3.1 Symptomatik (Spontanangaben der Patientin)

Die Klientin klagte über Ängste, innere Unruhe und Schlaflosigkeit. Außerdem habe sie häufig Kopf- und Halsschmerzen und Schmerzen im Rücken und in den Armen. Sie befinde sich in neurologischer Behandlung, es gebe jedoch keinen organischen Befund für die Schmerzen; der Arzt habe ihr eine Psychotherapie empfohlen und Antidepressiva verschrieben, die sie aber nicht nehmen wolle.

Die Klientin sagte, sie fühle sich insgesamt beruflich und familiär stark belastet; sie sei häufig sehr bedrückt und niedergeschlagen, sie empfinde ihre Partnerschaft als unbefriedigend; allein ihr Sohn bereite ihr Freude.

Sie bezeichnete sich selber als „eine Meisterin im Verdrängen“, darum falle es ihr schwer, über persönliche Dinge und ihre Gefühle zu reden.

3.2 Biographische und persönlichkeitsbezogene Daten

Die Klientin wuchs als jüngstes von drei Kindern auf, ein Bruder ist 5 Jahre älter, die Schwester 7 Jahre älter als die Klientin. Der Vater (vor einigen Jahren verstorben) war Arbeiter, die Mutter (heute 70jährig) Hausfrau. Der Vater wird von der Klientin als launisch und jähzornig beschrieben: Er sei Alkoholiker gewesen; er habe sie oft angeschrien und auch geschlagen.

Negativer noch als über den Vater fielen die Urteile über die Mutter aus - die, so wiederholte die Klientin mit Nachdruck, eine nervöse, ja gefühlskalte Frau gewesen sei. Ihrer Tochter habe sie mehrfach deutlich zu verstehen gegeben, daß sie ein unerwünschtes Kind gewesen sei. Das Gefühl, von ihrer Mutter geliebt zu werden, habe die Klientin nie gekannt.

Die Eltern hätten sich oft gestritten und so gut wie nie Zeit für die Kinder gehabt. Gemeinsame Unternehmungen habe es nicht gegeben.

Die Klientin war Linkshänderin und wurde mit Beginn der Einschulung um- erzogen. Die Lehrerin habe ihr mit dem Rohrstock auf die Finger geschlagen, wenn sie die linke Hand benutzt habe. Die Mutter habe sie zwar nicht geschlagen, aber sie ständig zur Rechtshändigkeit ermahnt und ihr alle Dinge in die rechte Hand gegeben.

Während der Pubertät sei ihr wenig erlaubt worden. Sie habe immer früh zu Hause sein müssen. Der Enge der Familie habe sie sich entzogen, indem sie häufiger für längere Zeit bei einer Freundin geblieben sei. „Wenn ich es nicht mehr aushielt, bin ich einfach abgehauen.“ Während der Lehre zur Kaufmännischen Angestellten sei es ihr oft schlecht gegangen, sie habe den Beruf nie erlernen wollen. „Hol‘ tief Luft und beiß‘ die Zähne zusammen!“, sei der Rat ihrer Mutter gewesen: „Sei froh, daß du überhaupt einen Beruf erlernst.“

Später, so berichtete die Klientin, habe sie auf dem zweiten Bildungsweg das Abitur nachgeholt, sie habe ein Studium begonnen, es aber wieder abgebrochen.

Die Klientin berichtete weiter, daß sie seit 17 Jahren unverheiratet mit ihrem Partner zusammenlebe; sie hätten einen gemeinsamen Sohn, der heute 7 Jahre alt sei. Sie selber und ihr Partner arbeiteten als Kaufmännische Angestellte in einem Betrieb, der dem Vater des Partners gehöre; auch der Bruder des Partners arbeite dort.

Beide - die Klientin und ihr Partner - seien sehr unzufrieden mit der beruflichen Situation. Die Verquickung von Berufs- und Familieninteressen führe ständig zu Spannungen in ihrer Beziehung: Der Partner sei oft krank, trinke viel und jammere, ändere aber nichts; er fühle sich für nichts verantwortlich; sie müsse sich um alles kümmern; wenn der Partner krank sei, müsse sie auch ihn noch betreuen.

Die Klientin erwähnte, jeder von ihnen (sie selber, aber auch ihr Partner) habe während der 17 Jahre ihres Zusammenlebens immer wieder Affären mit anderen gehabt; sie betonte jedoch, daß sie ihren Partner „trotz allem“ sehr gern habe.

Ihr Sohn, so berichtete die Klientin, entwickle sich gut, spüre aber mittlerweile auch die Spannungen zwischen seinen Eltern. Er mache noch ab und zu ins Bett. „Wahrscheinlich bin ich oft zu hektisch und ungeduldig mit ihm“, lautete die Ursachenzuschreibung der Klientin.

3.3 Diagnostische Befunde

Die Klientin wirkte im Erstgespräch sehr nervös und redete viel. Sie psychologisierte und rationalisierte ohne größere emotionale Beteiligung. Wurde sie aufgefordert, sich auf eine deutliche Aussage festzulegen, wich sie aus. Die Klientin zeigte einen rasch wechselnden Attributionsstil: Sie äußerte zunächst deutliche Selbstanklagen, etwas später jedoch sah sie die Ursachen für alle Probleme in ihrem sozialen Umfeld, vor allem in der beruflichen Situation und in der Partnerschaft. Andererseits gelang es ihr bei bestimmten Themen, innezuhalten und sich in konstruktiverer Weise selbstkritisch zu betrachten, z.B. in bezug auf die Beziehung zu ihrem Sohn, wobei eine stärkere innere Beteiligung spürbar wurde.

Insgesamt litt die Klientin unter deutlich gedruckter Stimmung bei einer gleichzeitig gesteigerten inneren Unruhe. Sie fühlte sich den Belastungen des Alltags nicht mehr gewachsen und äußerte Zukunftsängste. Sie zeigte nur noch geringes Interesse an Dingen, die ihr vormals Freude machten.

Angesichts der belastenden Lebenssituation und der körperlichen Beschwerden empfand die Klientin einen hohen Leidensdruck. Klar formulierte sie ihren Willen, daran etwas zu ändern. Die hohe Motivation und Einsichtsfähigkeit zusammen mit ihrer „kämpferischen“ Grundhaltung ließen einen guten Therapieerfolg erwarten. Die Klientin war bis dahin nicht in therapeutischer Behandlung.

4. Diagnose mit ICD - Nummer (IO. Revision)¹

F 32.11: mittelgradige depressive Episode mit somatischem Syndrom

5. Angaben zur Genese und Aufrechterhaltung der Störungen/Symptomatik

Die familiäre Situation, in der die Klientin aufwuchs, war geprägt durch die zum Teil offene Ablehnung von seiten der Mutter und insgesamt durch fehlenden Rückhalt und emotionale Zuwendung in der Familie. Hinzu kamen traumatisierende Erfahrungen in der frühen Schulzeit (gewaltsame Umerziehung zur Rechtshänderin).

Viel spricht dafür, daß die Klientin -um sich vor den Verletzungen zu schützen - in dieser Atmosphäre emotionaler Kälte und massiver Kränkungen eine starke Verslossenheit und einen hohen Grad an Mißtrauen gegenüber menschlichen Beziehungen entwickelte.

¹ ICD 10: „Internationale Klassifikation psychischer Störungen“ der Weltgesundheitsorganisation. Dilling et al. (1994).

Das elterliche Verhalten bei der Erziehung kann sowohl mit der depressiven Symptomatik der Klientin als auch mit ihrem inkonsistenten Attributionsstil in Zusammenhang gebracht werden: Das Verhältnis zum Vater war aus Sicht der Klientin bestimmt durch dessen unberechenbares Verhalten. Es war für die Klientin nicht vorherzusehen oder nachzuvollziehen, welche Reaktion in welcher Situation von ihm zu erwarten war.

Die Mutter war für die Klientin zwar zuverlässiger und präsenter, wurde aber trotz räumlicher Nähe emotional distanziert wahrgenommen. In diesem Sinne fühlte sich die Klientin von der Mutter besonders alleingelassen.

Auch mit den beiden deutlich älteren Geschwistern scheint die Klientin wenig Gemeinsamkeit erlebt zu haben. Sie hat früh lernen müssen, die Dinge mit sich selbst auszumachen. Daß andere Menschen ihr halfen und sie unterstützten, hat sie der eigenen Erinnerung nach bewußt zum ersten Mal in der Begegnung mit ihrer Freundin während der Pubertät wahrgenommen.

In ihrer eigenen Partnerschaft wiederholt die Klientin teilweise die Verhaltensmuster, die sie bei ihrer Mutter erlebt, aber abgelehnt hat. So übernimmt sie trotz einer emotionalen und auch körperlichen Distanz zu ihrem Partner eine fürsorgliche und „kümmernde“ Rolle. Andererseits begegnet sie den Problemen des Partners mit einer Haltung, die der der Mutter ihr gegenüber gleicht: „Der soll aufhören zu jammern und sich ‘mal zusammenreißen.“

6. *Behandlungsverlauf*

6.1. *Schilderung des bisherigen Verlaufs*

Zunächst war es wichtig, das Vertrauen der Klientin zu gewinnen und eine tragfähige therapeutische Beziehung aufzubauen. Eine gewährende und unterstützende therapeutische Haltung wirkte der immer wieder spürbaren evasiven Tendenz der Klientin entgegen und ermöglichte es ihr, sich zunehmend auf die therapeutische Situation einzulassen. Neben der anamnestischen Erhebung zielte die Therapie zunächst darauf, die aktuelle Situation - Belastung in Familie und im Beruf - zu besprechen und zu klären. Dabei stand allerdings zunächst die deutliche Tendenz der Klientin, über ihre Gefühle und Empfindungen „hinwegzureden“, einer tieferen Auseinandersetzung im Weg. Außerdem schwankte sie zwischen massiven Selbstanklagen und genauso heftigen Schuldzuweisungen an ihre Mitwelt, ohne daß eine deutliche Gefühlsbeteiligung spürbar wurde.

Im Verlauf des therapeutischen Prozesses gelang es der Klientin zunehmend, auf extrapunitive Deutungsmuster bzw. lähmende Selbstanklagen zu verzichten und mehr Möglichkeiten zu entdecken, auf die belastende Situation selbst einzuwirken und sie positiv zu verändern.

Der Klientin wurden eigene Verhaltensweisen verständlicher, wenn sie ihre Reaktionen in einen biographischen Zusammenhang einordnen konnte. Bei dieser konfliktzentrierten Arbeit wurden Rollenspiele bzw. die gestalttherapeutische Dialogtechnik („Leerer Stuhl“)² eingesetzt. Zur diagnostischen Klärung der familiären Beziehungen wurde eine Familienskulptur erstellt.

Im geplanten weiteren Verlauf der Behandlung sollen neben dem konfliktzentrierten therapeutischen Gespräch weiterhin erlebnisaktivierende gestalttherapeutische oder psychodramatische Techniken eingesetzt werden.

6.2 Veränderung der Beschwerden bzw. der Symptomatik

Die Klientin ist zunehmend in der Lage, ihre Problematik in einen lebensgeschichtlichen Zusammenhang zu stellen. Ihre aktuelle Situation erscheint ihr nicht mehr durch ein ungünstiges Umfeld verursacht, sondern durch eigenes Zutun veränderbar. Das betrifft auch die körperlichen Beschwerden, die seit Beginn der Therapie deutlich nachgelassen haben.

Die Rationalisierungstendenzen, die eine tiefergehende Auseinandersetzung mit ihrer Problematik verhindert haben, hat sie allmählich aufgegeben. Sie ist emotional deutlich beteiligt und läßt ansatzweise auch sehr schmerzhaft Gefühle zu. Sie entdeckt zunehmend mehr Möglichkeiten, ihre Wünsche und Bedürfnisse zu erkennen und auszudrücken, insbesondere auch bzgl. ihrer Partnerschaft.

6.3 Beschreibung der Symptome und Probleme zum jetzigen Zeitpunkt

Die Symptome und ihr psychischer Hintergrund schränken die Klientin in ihrer Lebensführung weiterhin ein. Partnerschaft und aktuelle Berufsarbeit werden weiterhin als sehr belastend erlebt, auch wenn die Klientin sich bemüht, das in der Therapie Gelernte in ihrer sozialen Umgebung umzusetzen.

Die Belastung hat sich zum gegenwärtigen Zeitpunkt verschärft. Der Grund liegt darin, daß der Bruder des Partners den Betrieb übernommen hat und die Spannungen bei der Arbeit in einem Maße zugenommen haben, daß beide - Klientin und Partner - kündigen wollen.

² „Leerer Stuhl“: gestalttherapeutisches Verfahren, bei dem ein Klient einen (nicht tatsächlich anwesenden) Dialogpartner auf einen leeren Stuhl „setzt“. Im Gespräch nimmt er selbst dessen Rolle ein und wechselt so ständig zwischen seinem ursprünglichen Platz und dem „leeren Stuhl“. Dialogpartner können andere Personen, „Teilpersönlichkeiten“ (z.B. Ja-Sage/Nein-Sager), Tiere, Vorstellungen oder auch Dinge sein, die im Alltagsleben oder in Träumen für den Klienten von Bedeutung sind. Die Methode dient der Aktualisierung, Konkretisierung und unmittelbaren Verarbeitung belastender Erlebnisse (Walter, 1996).

³ Familienskulptur: familientherapeutisches Verfahren, bei dem ein Klient Familienmitglieder (tatsächlich anwesende Familienmitglieder oder - in der Einzeltherapie - symbolisiert durch Stühle, Kissen, Puppen o.ä.) im Sinne von wahrgenommener emotionaler Nähe oder Distanz um die eigene Person herum gruppiert. Die Methode dient dazu, die Position des Klienten in der Familie anschaulich darzustellen. Durch Rollentausch kann der Klient auch Positionen anderer Familienmitglieder einnehmen und deren Blickwinkel erfahren. Das Verfahren dient sowohl diagnostischen als auch interventiven Zwecken (Cierpka, 1988).

7. Therapieplanung

7.1 Beschreibung der weiteren therapeutischen Maßnahmen

Die Klientin ist stark motiviert, die Behandlung fortzusetzen und arbeitet auch außerhalb der Therapie aktiv an der Änderung ihres Verhaltens. Sie kann zunehmend ihrem emotionalen Erleben Ausdruck verleihen.

Sie stellt sich mittlerweile auch sehr unangenehmen und schmerzhaften Gefühlen. Diese Veränderung ihrer Erlebnisfähigkeit wird von der Klientin aber begrüßt; sie hat das Gefühl, auf dem richtigen Weg zu sein.

Das vordringliche Ziel der Therapie ist es, die Erlebnis- und Beziehungsfähigkeit der Klientin zu steigern und ihre Ausdrucksmöglichkeiten zu verbessern.

7.2 Begründung der Notwendigkeit der Therapieverlängerung

Der bisherige Therapieverlauf hat zu einer stabilen therapeutischen Beziehung geführt. Die Klientin konnte wichtige Einsichten bezüglich ihrer Problematik gewinnen und ihren Alltag merklich verändern. Die depressive Reaktion konnte gemildert werden. Zunehmend wird es ihr möglich, sich mit angstbesetzten Situationen auseinanderzusetzen.

Eine tiefere Auseinandersetzung mit diesen schmerzhaften Erlebnissen hat allerdings erst in den letzten Therapiesitzungen begonnen. Der Zustand der Klientin ist in dieser Hinsicht zur Zeit noch instabil. Sie bedarf der weiteren therapeutischen Unterstützung.

7.3 Prognostische Einschätzung

Die Prognose für den weiteren Therapieverlauf kann als günstig eingeschätzt werden. Die Klientin hat eine hohe Motivation, da sie einerseits immer noch einen deutlichen Leidensdruck verspürt, andererseits aber auch gute Fortschritte im bisherigen Therapieverlauf gemacht hat.

8. Zusammenfassung

Die Klientin leidet unter einer mittelgradigen depressiven Episode, die sich vor allem in Ängsten und innerer Unruhe äußert, begleitet von psychosomatischen Beschwerden. Die zugrundeliegende Psychodynamik konnte in ersten Ansätzen für die Klientin offengelegt werden. Aktuelle Probleme wurden bearbeitet. Nach gewonnener Stabilität der therapeutischen Beziehung ist die Klientin zunehmend in der Lage, sich angstbesetzten Situationen zu stellen. Dieser Schritt bedarf der weiteren therapeutischen Unterstützung.

21. Kapitel

Beispiel II Integrativer Diagnostik

Psychologische Begutachtung

Kapitel 21 bringt ein Beispiel integrativer Diagnostik, das von der diagnostischen Aufgabe her eher der *Klassifikation* zugehört: die Psychologische Begutachtung.

Integration bezeichnet (es sei wiederholt) ein diagnostisches Urteil, das auf Informationen beruht, die multimodal gewonnen werden, das heißt mit so unterschiedlichen Verfahrensklassen wie Leistungs- und Persönlichkeitstests, mit Exploration oder projektiven Verfahren.

Wir bilden fünf Teilabschnitte:

- Abgrenzungen (Definitionen) (21.1),
- Psychologische Begutachtung im sozial-ethischen Kontext (21.2),
- Gutachten-Gliederung: Überblick (21.3),
- Gutachten-Gliederung: Darstellung der einzelnen Abschnitte (21.4),
- Fehlertendenzen (21.5).

Es folgen eine Zusammenfassung (21.6) und eine Reihe von Kontrollfragen (21.7).

21.1 Abgrenzungen (Definitionen)

Psychologische Gutachten werden auf vielen Tätigkeitsfeldern und zu vielen Fragestellungen angefordert.

Beispiele für diese Vielfalt geben die „Richtlinien für die Erstellung Psychologischer Gutachten“ (dpv, 1994 b, 12-16):

- So lassen **Schulen** Lernfähigkeit, Verhaltensauffälligkeiten, Wahl von Schulformen beurteilen usw.
- **Versicherungsträger** lassen Gutachten zu Problemen wie Rentenfragen, Berufsunfähigkeit, psychotherapeutischen Interventionen usw. erstellen.
- **Strafgerichte** fordern Gutachten an zu Schuldfähigkeit, Glaubwürdigkeit von Zeugenaussagen, Zumutbarkeit normgerechten Verhaltens (Putativnotstand) usw.

- **Zivilgerichte** verlangen Gutachten zu Fragen wie Prozeßfähigkeit, Delikt-haftung, Schadenersatz- oder Schmerzensgeldforderungen usw.
- **Familiengerichte** stützen sehr oft ihre Entscheidungen auf Gutachten, wenn sie über das elterliche Sorgerecht zu befinden haben.

Können sich Gutachten so unterschiedlicher Tätigkeitsfelder an einem einheitlichen Erstellungsschema orientieren? Das zu erwarten wäre unangemessen. Dennoch gibt es „Standards und Kriterien“, an denen sich alle Arten von Gutachten messen müssen, so die „Richtlinien für die Erstellung Psychologi-scher Gutachten“ (dpv, 1994 b).

Wenn wir in diesem Kapitel ein bestimmtes Arbeitsschema vorstellen, dann nur aus der Absicht, dem Anfänger eine Möglichkeit der Einübung zu veran-schaulichen.

Was die Konzeption psychologischer Gutachten angeht, so bieten die „Richt-linien“ (dpv, 1994, 8-9) folgende Umschreibung (Definition) an:

„Ein ... Psychologisches Gutachten ist eine wissenschaftliche Leistung, die darin besteht, aufgrund wissenschaftlich anerkannter Theorien und Kriterien nach feststehenden Regeln der Gewinnung und Interpretation von Daten zu konkreten Fragestellungen Aussagen zu machen. Es handelt sich um die Ant-wort eines Experten, des Diplom-Psychologen, auf Fragen, zu denen er auf-grund seines Fachwissens, des aktuellen Forschungsstandes und seiner Erfah-rung Stellung nimmt. Ein solches Gutachten muß umfassen:

- die Fragestellung,
- die Untersuchungsverfahren,
- die relevanten Daten,
- deren Interpretation und
- die Schlußfolgerungen des Gutachters.

Im Dienste einer klaren Sprachregelung sollten darum Leistungen wie Psy-chologische Stellungnahmen, Gutachterliche Stellungnahmen oder Untersu-chungsbefunde nicht als Psychologische Gutachten bezeichnet werden. Diese Begriffe beschreiben ihrerseits wichtige Teilbereiche gutachterlicher Tätigkeit:

- Als **„Gutachterliche Stellungnahme“** wird die psychologische Antwort auf eine eingeschränkte Einzelfrage bezeichnet.
- Als **„Psychologische Stellungnahme“** wird die Stellungnahme zu einem Gutachten oder einer Fragestellung ohne eigene Befunderhebung bezeich-net.
- Als **„Untersuchungsbefund“** wird eine für Nicht-Psychologen verständlich aufbereitete Aussage über Ergebnisse einer Untersuchung mit psychologi-schen Verfahren bezeichnet.“

Der Prozeß der Begutachtung läßt sich in drei Abschnitte gliedern:

1. Annahme und Klärung der gutachterlichen Fragestellung in ihrem Kontext,
2. Datenerhebung zur Gutachtenerstellung,
3. Integration der Daten zu einem Gutachtentext.

Dem Prozeß der Begutachtung folgt sehr oft ein Prozeß der Intervention; denn die Stellungnahme in einem Gutachten resultiert in der Regel in einen Handlungsvorschlag, der eine Intervention anzielt. Damit ist die Frage nach prognostischen Aussagen angeschnitten.

über den Abschnitt (1: Gutachten-Annahme) und (2: Datenerhebung) gibt Kapitel 19 Auskunft, wenigstens in allgemeiner Form (vgl. S. 413). Darum soll dieses Kapitel 21 vor allem auf den Abschnitt (3) eingehen: auf die Integration der Daten zu einem Gutachtentext.

Zur Gestalt des Gutachtens

Was die Gestalt des Gutachtens angeht, so orientieren wir uns an dem Vorschlag von Thomae, der das psychologische Gutachten als einen Versuch der Kommunikation zwischen Experten und Laien konzipiert. Mit Laie ist der Fragesteller gemeint, mit Experte der Diplompsychologe als Gutachter (Thomae, 1967, 745).

21.2 Psychologische Begutachtung im sozial-ethischen Kontext

Psychologische Begutachtung vollzieht sich (wie jeder andere diagnostische Akt) in einem ethisch-juristischen Kontext. Den mehrschichtigen Zusammenhang haben viele Autoren analysiert, etwa Amelang und Zielinski (1994), Arnold (1972), Hartmann (1984), Haubl (1984), Heiß (1964), Jäger, R. S. (1986), Keßler (1982), Kipnowski (1981), Kubinger (1995 c), Pulver, Lang und Schmid (1978), Schmidt, L.R. (1982, 1995), Westhoff und Kluck (1991).

Haubl hat eine Reihe von Empfehlungen formuliert, welche die komplexen, auch disparaten Zusammenhänge aufzeigen, in die das Psychologische Gutachten einbezogen ist (1984, 73-74).

21.3 Gutachten-Gliederung: *Überblick*

Wenn das Gutachten als Text Rechenschaft geben soll über die Begutachtung als Prozeß, dann empfiehlt sich folgende Gliederung:

- In einem ersten Abschnitt umreißt der Gutachter sein Untersuchungsdesign: Er formuliert die diagnostische Frage, er nennt Fragesteller, Untersucher und Adressaten, er nennt die Untersuchungstermine und gibt die Datenquellen an. Dieser Abschnitt läßt sich als **Übersicht** bezeichnen.
- In einem zweiten Abschnitt listet er Informationen auf, die er vorfindet, wenn er seine Untersuchung beginnt. Er führt alle Quellen an, aus denen er geschöpft hat. Dieser Abschnitt heißt **Vorgeschichte**.

- In einem weiteren Abschnitt, genannt **Untersuchungsbericht**, stellt der Gutachter die Informationen zusammen, die er selbst beim Probanden erhoben hat.
- In einem vierten Abschnitt faßt er die Informationsvielfalt aus Vorgeschichte und Untersuchungsbericht thematisch zusammen, er integriert Einzelheiten zu übergreifenden Aussagen. Dieser Teil heißt **Befund**.
- In einem letzten Abschnitt formuliert er seine Antwort als **Stellungnahme**. Er kann eine Diagnose und eine Prognose abgeben, kann sich aber auch auf einen Entscheidungsvorschlag beschränken,

21.4 Gutachten-Gliederung:

Darstellung der einzelnen Abschnitte

Die Funktion der einzelnen Teile sei nun erklärt und an Beispielen veranschaulicht:

- Übersicht (21.4.1),
- Vorgeschichte (21.4.2),
- Untersuchungsbericht (21.4.3),
- Befund (21.4.4),
- Stellungnahme (21.4.5).

Vorbemerkung zu den Fallbeispielen

Alle gutachterlichen Beispiele, die in Kasten 21-1 bis 21-10 aufgeführt werden, stammen aus „echten“ Fällen. Darum konnte ich mich nicht entschließen, bei dieser Neuauflage andere Verfahren ‚einzusetzen‘: neuere Diagnostica anstelle der älteren. Doch habe ich die Liste der Verfahren gekürzt, wo es möglich war.

Dieses Vorgehen sehe ich insofern berechtigt, als dieses Kapitel das Gutachten vor allem unter dem Aspekt der formalen Gestaltung behandelt, nicht unter dem der Bekanntgabe neuer Verfahren.

Aus Gründen des Datenschutzes wurden Name, Beruf und Wohnort der betroffenen Personen so geändert, daß die Anonymität hinreichend gewahrt wurde.

21.4.1 Erster Abschnitt des Gutachtens: *Übersicht*

Die Übersicht soll den Leser in ähnlicher Weise informieren wie Titel und Inhaltsverzeichnis eines Buches. Sie benennt das Problem, sie gibt an, wer es gestellt und wer es bearbeitet hat, sie nennt Untersuchungsinstrumente und Untersuchungstermine.

Darum empfehlen sich einige **Sprachregelungen**. Die Hauptangaben lassen sich in Form eines Briefkopfes anordnen. Darin erscheinen

- der Gutachter (und seine Mitgutachter),
- der Adressat als Empfänger,
- ein ‚Betreff‘, das stichwortartig die Fragestellung umreißt,
- ein ‚Bezug‘, falls gegeben, mit den entsprechenden schriftlichen und mündlichen ‚Vorgängen‘ (Briefen, Telefonaten),
- eine Anrede, wenn eine bestimmte Person oder Personengruppe der Adressat ist, oder aber
- eine neutrale Angabe „Gutachten gemäß...“.
- ein Abschnitt, der in wenigen Sätzen die Aufgabe des Gutachtens erklärt,
- die Auflistung der Quellen, auf die sich das Gutachten stützt.

Die Übersicht schließt **nicht** mit einer Formel nach Art eines Briefschlusses, etwa mit einer **Unterschrift**. (Die Unterschrift leistet der Gutachter am Ende des Gutachtens, also nach der Stellungnahme.)

Ein Beispiel soll Aufbau und Rolle der Übersicht verdeutlichen (Kasten 21-1).

Kasten 21-1:
Zum ersten Abschnitt eines Gutachtens: ‚Übersicht‘

Psychologische Praxis
Edmund R., Diplom-Psychologe
Riener Straße 181
55431 Kalenberg

30. Juni 1987

An das
Familiengericht am Amtsgericht K.
z. Hdn. Herrn Richter J.
Postfach 873
55431 Krallenhofen

Betreff: Familienrechtssache L.
Hier: Tobias L., geboren am 13.3.1977

Bezug: Schreiben des Amtsgerichts vom 24.4.1987
Geschäfts-Nr. 7 XX 88/26

Gutachten

nach Beschluß des Amtsgerichts vom 22.4.1987 und Schreiben des Amtsgerichts vom **24.4.1987**.

Gemäß dem genannten Beschluß vom 22.4.1987 soll das Gutachten klären, ob „bei der augenblicklichen psychischen Verfassung des jetzt zehnjährige Tobias L. eine Trennung von der Mutter zu Beginn des neuen Schuljahres mit hoher Wahrscheinlichkeit zu schwerwiegenden Komplikationen führt“. Der Vater des Jungen hat eine solche Trennung beantragt, er schlägt eine Internaterziehung für Tobias vor.

Das Gutachten stützt sich auf

- das Studium der einschlägigen **Akten**,
- eine **Unterredung** mit dem **Kindesvater**, Herrn Diplom-Mathematiker Manfred L., am 12. Mai 1987,
- eine **Unterredung** mit der **Kindesmutter**, Frau Helma L., am 19. Mai 1987, sowie

- eine **Untersuchung** von **Tobias L.**, die am 21. Mai 1987 in meiner Praxis stattfand, bei der folgende Verfahren einbezogen wurden:
 - eine ausführliche Exploration mit Tobias L. durch meine Mitarbeiterin, Frau Diplom-Psychologin Monika M.,
 - der Grundintelligenztest, Skala 2, von Cattell in der Bearbeitung von Weiss (CFT 2)
 - das Prüfsystem für Schul- und Bildungsberatung von Horn (PSB),
 - usw.

Kommentar zu Kasten 21-1: Die Liste der Verfahren wurde gekürzt. - Wie aus den Akten hervorgeht, hat sich das Ehepaar L. getrennt (ohne sich scheiden zu lassen). Tobias ist bei der Mutter geblieben. Dem Vater ist die Erziehung der Mutter jedoch zu ‚weich‘, die Schulleistungen des Jungen bleiben hinter den väterlichen Erwartungen zurück. Er wendet sich an das Familiengericht: Ihm soll das Recht übertragen werden, die Schullaufbahn seines Sohnes allein zu bestimmen; er will durchsetzen, daß Tobias in ein Internat geht.

Aus dem Schreiben des Amtsgerichts wird die Fragestellung als Zitat übernommen (ob „bei der augenblicklichen psychischen Verfassung des jetzt zehnjährigen Tobias L. eine Trennung von der Mutter zu Beginn des neuen Schuljahres mit hoher Wahrscheinlichkeit zu schwerwiegenden Komplikationen führt“). Der anschließende Satz im Briefkörper kommentiert, was ‚Trennung von der Mutter‘ bedeutet, nämlich Internatserziehung. Eine ausführliche Hintergrundinformation gehört in die Vorgeschichte oder in den Untersuchungsbericht (Anamnese oder Exploration).

Die Auflistung der Quellen gliedert sich nach Verfahrensklassen: Akten, Gespräche, Leistungstests usw.

Das Beispiel wird im nächsten Abschnitt, der Vorgeschichte, fortgeführt.

21.4.2 Zweiter Abschnitt des Gutachtens: ***Vorgeschichte***

Die Vorgeschichte übernimmt im Gutachten jene Funktion, die bei einer wissenschaftlichen Arbeit der Literatursichtung zufällt. Es soll geklärt werden, was über die Fragestellung schon an Informationen vorliegt. Es geht um den **Entdeckungszusammenhang** der diagnostischen Frage.

Hilfreich kann jede Informationsquelle sein, die Licht auf die Fragestellung wirft: Akten von Verwaltungen oder Gerichten, Zeugnisse von Schule oder Lehrstellen, medizinische Atteste, auch umfassende psychologische Vorgutachten.

Vorinformationen zu sichten zielt darauf, den ‚Entdeckungszusammenhang‘ der diagnostischen Frage ernstzunehmen, also wichtige Aspekte zu berücksichtigen, zu denen schon Informationen vorliegen. Umgekehrt ginge ein Untersucher vor, der sich nur auf aktuelle Aussagen des Probanden stützen wollte: Er nähme das Risiko in Kauf, daß der Proband ‚aus wohlverstandenen Eigennutz‘ Informationen verschweigt oder ‚vergißt‘, die für die Frage von Belang sind.

Weil die Vorgeschichte eine Quellensammlung bietet, sollte ihre Abfassung **Sprachregelungen** folgen, die diesen Charakter kenntlich machen:

- Die *Informationsquelle* wird eindeutig angegeben. (An keiner Stelle der Vorgeschichte darf unklar bleiben, woher eine Information stammt.)
- Wenn Angaben einer Quelle zusammengefaßt werden, sollte zur Kennzeichnung der mittelbaren Wiedergabe die Aussage in den *Konjunktiv* treten. (Angaben wie etwa Geburt, Geburtsort, Beruf usw. können auch im Indikativ erscheinen.)
- Die Vorinformationen sollen so ausgewählt werden, daß die Fragestellung in ihren *Kontext* eingebettet erscheint, die Darstellung aber nicht ungebührlich anschwillt.

An dem letzten Satz wird deutlich, daß Begutachtung hier subjektive Freiheitsgrade in großem Umfange gewährt. Tönung der Wiedergabe, Umfang der referierten Quellen, Wertungen, mit denen sie eingeleitet werden: Alle diese Gestaltungsschritte liegen im Ermessen des Gutachters. Darum die Dringlichkeit einer Orientierung an der Fragestellung!

Hingewiesen sei auf eine besondere **Gefahr**: Liegen zu einem Fall umfangreiche Akten(notizen) vor, dann kann es geschehen, daß irrelevante Informationen aufgenommen und gegebenenfalls von Gutachten zu Gutachten weitergeschleppt werden. Ungünstig für den Probanden muß diese Weitergabe sich auswirken, wenn es sich um negative Aussagen handelt. Die Schwierigkeiten, eine angemessene Kommunikation über die Sachverhalte eines Gutachtens herzustellen, hat Spitznagel ausführlich geschildert (1984 b): Die Vorgeschichte ist *eine zentrale „Stelle“*, an der Mißverständnisse sowohl erzeugt als auch weitervermittelt werden können.

Darum der Rat, die Vorgeschichte - bezogen auf die Fragestellung - möglichst knapp zu fassen und immer den Bezug zu den Informationsquellen sichtbar zu machen.

Es folgen **zwei Beispiele**:

- Das erste in Kasten 21-2 führt den Fall aus dem vorhergehenden Abschnitt (Fall Tobias L.) weiter,
- das zweite in Kasten 21-3 greift einen neuen Fall auf (der später, im Befund und bei der Stellungnahme, fortgeführt wird).

Kasten 21-2:

Zum zweiten Abschnitt eines Gutachtens: ‚Vorgeschichte‘

(Fragestellung: Unterbringung von Tobias L. in einem Internat:
vgl. Kasten 21-1, S. 443)

Den Akten ist zu entnehmen, daß bei Tobias L. seit längerer Zeit schulische Schwierigkeiten aufgetreten sind. In einem Gutachten vom 22. Februar 1984 führte die Psychologin des Kinderkrankenhauses in A., Frau Diplom-Psychologin Marianne H., die Probleme auf Legasthenie zurück.

In einem Gutachten des Diplom-Psychologen Walter W., datiert vom 30. März 1986, wurde die Diagnose ‚Legasthenie‘ bestätigt, zugleich aber die Vermutung geäußert, Tobias sei bei seiner Einschulung 1983 noch nicht schulreif gewesen, denn er habe die erste Schulklasse wiederholen müssen.

Gemäß den Akten hat der Vater, Herr Diplom-Mathematiker Manfred L., im März 1987 beantragt, daß ihm das Recht übertragen werde, die schulische Ausbildung von Tobias allein zu bestimmen. Anlaß für den Antrag sei gewesen, daß Herr L., um die Schulschwierigkeiten endgültig zu beheben, seinen Sohn in einem Internat habe unterbringen wollen, daß Frau L. diesem Vorhaben jedoch widersprochen habe. - Die Eltern sind nicht geschieden, leben aber getrennt.

Kommentar zu Kasten 21-2: Die Akten enthalten einen ausführlichen Briefwechsel zwischen den Anwälten der beiden Parteien, ebenso alle Schulzeugnisse von Tobias und mehrere psychologische Gutachten über ihn. Nur ein begrenzter Teil der Informationsfülle geht in dieses Beispiel ein.

Jeden der drei Abschnitte leitet ein Satz im Indikativ ein, es handelt sich um Bezugnahmen auf die Quellen (die Akten, die das Gericht zur Verfügung gestellt hat) oder um Angaben von Fakten.

Die Aussagen über die Vermutung des Psychologen (im zweiten Abschnitt) und über die Inhalte des väterlichen Antrages (im dritten Abschnitt) stehen im Konjunktiv.

Kasten 21-3:

Zum zweiten Abschnitt eines Gutachtens: ‚Vorgeschichte‘

(Fragestellung: Führerschein-Entzug wegen ‚Alkohols am Steuer‘.)

Herr V. erwarb, wie aus der Führerscheininakte hervorgeht, die Fahrerlaubnis (der Klasse 3) im Jahre 1960. Zweimal hat er die Fahrerlaubnis wieder verloren: 1974 wegen eines Unfalls, den er verursacht hatte, 1979 wegen fahrlässiger Straßenverkehrgefährdung. In beiden Fällen stand er unter Alkoholeinfluß.

Zu dem ersten Führerscheinentzug kam es, gemäß den Akten, folgendermaßen: Am 1. Juli 1974 habe Herr V. in W. einen entgegenkommenden Pkw gestreift. Die Blutalkoholkonzentration habe sich auf 2,23 Promille belaufen. Durch Beschluß des Amtsgerichts B. vom 9. Juli 1974 sei der Führerschein vorläufig eingezogen, durch Gerichtsurteil vom 18. Dezember 1974 sei Herr V. zu einer Geldstrafe von 600 DM verurteilt worden. Darüber hinaus sei ihm für weitere drei Monate die Fahrerlaubnis entzogen worden.

Am 30. April 1975 wurde Herrn V., so die Akten, die Fahrerlaubnis wieder erteilt.

Grund für den zweiten Entzug des Führerscheins war, laut den Akten, folgender Vorfall: Am 30. Januar 1979 habe Herr V. unter Alkoholeinfluß mit seinem Wagen eine Gartenmauer in M. beschädigt, ein dort parkendes Auto, einen Fiat 127, gerammt und dann Fahrerflucht begangen. Durch Strafbefehl des Amtsgerichts B. vom 12. Februar 1979 sei „wegen fahrlässiger Straßenverkehrgefährdung in Tateinheit mit Trunkenheit am Steuer und Fahrerflucht“ eine Geldstrafe von 700 DM verhängt und die Fahrerlaubnis entzogen worden mit einer Wiedererteilungs-Sperrfrist von sieben Monaten.

Am 12. September 1979 habe Herr V. einen Antrag auf Wiedererteilung der Fahrerlaubnis (der Klasse 3) gestellt. Die Wiedererteilung sei aber abhängig gemacht worden von dem Ergebnis einer medizinisch-psychologischen Untersuchung.

Diese Untersuchung hat am 28. und 29. September in . . . stattgefunden.

Kommentar zu Kasten 21-3: Hier werden den Akten nur die Angaben entnommen, die einen unmittelbaren Bezug zum Antrag auf Rückgabe des Führerscheins haben.

21.4.3 Dritter Abschnitt des Gutachtens: *Untersuchungsbericht*

Der Untersuchungsbericht gibt Rechenschaft, welche Informationen der Gutachter bei dem Probanden erhoben hat. Dabei können sehr viele Einzelinformationen zusammentreffen. Darum wird empfohlen, über die Einzelergebnisse in zwei Schritten zu berichten: in einem ersten Schritt die Einzelergebnisse auszubreiten, in einem zweiten die Einzelheiten nach Themengebieten zu integrieren. Die Einzelergebnisse stellt der Untersuchungsbericht vor, die Integration der nächste Abschnitt des Gutachtens, der Befund.

Gliederung des Untersuchungsberichtes

Um das diagnostische Urteil sorgfältig zu kontrollieren und nicht vorzeitig definitive Aussagen zu machen, sollte der Gutachter folgende **Gliederung** beachten:

- Er referiert über *jedes einzelne Verfahren in einem eigenen Bericht*. Als Grundriß für die Einzelberichte schlagen die „Richtlinien für die Erstellung Psychologischer Gutachten“ eine vierteilige Gliederung vor (Richtlinien dpv, 1994 b, 11):

1. Kurzbeschreibung der angewandten psychodiagnostischen Instrumente: **Testbeschreibung**;
2. Beschreibung der für die Fragestellung relevanten Verhaltensweisen des Probanden: **Verhaltensbeobachtung**;
3. Mitteilung der Ergebnisse, die für die Beantwortung der Fragestellung wichtig sind: **Ergebnisbericht** oder **Ergebnisteil**;
4. Interpretation der Ergebnisse nach den wissenschaftlich-psychologisch gegebenen Regeln (und, soweit notwendig, Hinweise auf Grenzen der Interpretierbarkeit der Daten): **Interpretation**.

Zu 1.: Die **Testbeschreibung** erklärt dem Laien als Empfänger, welche Informationen das einzelne Verfahren liefert, beispielsweise welche Gesamt- und Teilfunktionen - nach Meinung des Testautors - erfaßt werden, welche Bezugsnormen vorliegen (z. B. Normgruppen nach Alter, Geschlecht, Beruf). Gegebenenfalls sollte der Gutachter darauf hinweisen, daß die Eigenart eines Verfahrens dazu nötigt, die Ergebnisse zu relativieren.

Zu 2.: Die **Verhaltensbeschreibung** gibt an, in welcher Weise der Proband bei den einzelnen Verfahren mitgearbeitet hat.

Zu 3.: Der **Ergebnisbericht oder Ergebnisteil** legt dem Adressaten die wichtigsten Resultate vor. Aus Tests und Fragebogen referiert er

die numerischen Werte, nennt die zugehörigen Konstrukte und ihren Ausprägungsgrad.

Zu 4.: Die **Interpretation** sagt dem Empfänger als Laien, welche diagnostischen Aussagen im Ergebnisteil den Testwerten des Ergebnisberichtes enthalten sind.

- **Varianten:** Die vierteilige Gliederung des Untersuchungsberichtes sollte der Gutachter mit Blick auf Fragestellung und Empfänger so frei variieren, daß sie ihm eine Hilfe bietet - er sollte in ihr keine ‚Zwangsjacke‘ sehen:

Beispiele für Abwandlungen: Wenn der Empfänger ein Verfahren bereits kennt, entfällt die Testbeschreibung. - Eine Verhaltensbeschreibung kann, im Einzelfalle, überflüssig sein. - Bei manchen Verfahren, z.B. beim TAT, lassen sich die technischen Details einer Auswertung kaum im Ergebnisteil darstellen. - Wie breit eine Interpretation ausfällt, ob sie einen fortlaufenden Text bildet oder eine Liste von Stichworten bleibt, sollte der Gutachter fallbezogen entscheiden.

- Eine *Sonderstellung* nimmt der Untersuchungsbericht zu *Gesprächen* ein. Darüber wurde berichtet im Anschluß an die Rolle der Exploration für Diagnostik und Intervention (unter dem Stichwort der ‚Thematischen Zusammenfassung‘, S. 239). Die Abfassungsregeln seien kurz rekapituliert.

Untersuchungsbericht zu Anamnese, Exploration, Interview

Ein Untersuchungsbericht über Anamnese, Exploration oder Interview stellt die Schilderung des Probanden möglichst genau, aber in gestraffter und in thematisch gegliederter Form dar:

- Zu Beginn wird das *Erhalten* geschildert, das der Proband während des Gespräches gezeigt hat.
- Zur Kennzeichnung der mittelbaren Wiedergabe stehen die *Aussagen im Konjunktiv*. Unstrittige Angaben können auch im Indikativ erscheinen (z.B. Alter, Geburtsort).
- Biographische Angaben werden *chronologisch* berichtet, Aussagen über die augenblickliche Situation werden *zu thematischen Einheiten* zusammengezogen.
- Themen, die der Proband in seiner Schilderung *hervorhebt*, sollten eigens gekennzeichnet werden.
- *Zitate* sollte man *sparsam* verwenden. Sie erscheinen nur dann, wenn sie eine Einstellung oder ein Verhalten des Probanden besonders treffend veranschaulichen.
- „Herabsetzende und verletzende Ausdrücke müssen vermieden werden, sofern es sich dabei nicht um direkte Rede untersuchter Personen handelt“ (Richtlinien dpv, 1994 b, 7). Doch muß der Gutachter darauf achten, die Zitate im Kontext zu belassen. *Zitate, aus dem Zusammenhang gerissen, können erheblichen Schaden anrichten!*
- Eine *Interpretation* der Aussagen ist ein zusätzlicher Auswertungsschritt, darum auch in der Darstellung zu trennen von der Wiedergabe des Gespräches, indem man sie zum Beispiel als Zusammenfassung an den Schluß des Untersuchungsberichtes setzt.

Zwei Sprachregelungen für den Untersuchungsbericht

Die Deutung des Untersuchungsberichtes, wie sie hier vorgeschlagen wird, legt zwei Sprachregelungen nahe:

1. *Verhaltensbeschreibung*, Ergebnisbericht und Interpretation werden *im Imperfekt* abgefaßt.
2. Die *Interpretation* wird *unpersönlich* formuliert.

zu 1.: Verhaltensbeschreibung, Ergebnisbericht und Interpretation **im Imperfekt** zu berichten, dieser Rat zielt auf ein bestimmtes Anliegen: Beobachtungen und Ergebnisse des einzelnen Untersuchungsberichtes sollen dem Einzelverfahren und der Einzelsituation zugeordnet bleiben, sie sollen als **test-** und **situationsbezogene Aussagen** betrachtet und noch nicht als definitive Aussagen über den Probanden gewertet werden.

zu 2.: Dem gleichen Ziel dient der Vorschlag, die Interpretation **unpersönlich** zu formulieren.

Beispiel: *Herr X habe in einem Fragebogen, der Aggression erfaßt, den Staninewert 8 erhalten.*

Die **persönliche Prädikation** würde lauten: „Herr X ist überdurchschnittlich aggressiv.“

Die **unpersönliche Formulierung** lautet: „Ein Staninewert von 8 spricht für eine überdurchschnittliche Aggressivität.“

Die unpersönliche Prädikation identifiziert im Untersuchungsbericht noch nicht jedes einzelne Merkmal mit der Persönlichkeit des Probanden. Sie schafft Spielraum dafür, daß auch gegensätzliche Aussagen zugelassen und im Befund diskutiert werden.

Resümee zum Untersuchungsbericht: Bevor Beispiele geboten werden, seien die Abfassungsregeln zusammengefaßt:

- Der Untersuchungsbericht referiert die *Ergebnisse*, die der *Gutachter* beim Probanden erhoben hat.
- Ein Untersuchungsbericht wird *für jedes Einzelverfahren* erstellt.
- In der Regel sollte sich dieser Bericht *in vier Abschnitte gliedern*: Testbeschreibung, Verhaltensbeschreibung, Ergebnisbericht, Interpretation.
- Verhaltensbeobachtung, Ergebnisbericht und Interpretation werden *im Imperfekt* referiert.
- Die Interpretation sollte *unpersönlich* gefaßt werden.
- Exploration, Anamnese, Gespräche werden nicht wörtlich wiedergegeben, sondern *thematisch gegliedert und zusammengefaßt*. Zur Kennzeichnung der mittelbaren Wiedergabe sollten die Aussagen des Probanden *im Konjunktiv* stehen. Zitate sollte man sparsam verwenden.

Drei Auswertungsschritte

Bei Verfahren, die eine Vielzahl von Einzelergebnissen erbringen, lohnt es sich, den Untersuchungsbericht in drei Schritten zu erstellen:

Schritt 1: Ein Verfahren wird *ausgewertet* nach Zahlen und Kürzeln.

Schritt 2: Die Zahlen und Kürzel werden in Stichworten *interpretiert*.

Schritt 3: Es wird ein *fortlaufender Text* abgefaßt.

Schritt 1, 2 und 3 wiederholen in Kürze, was Kapitel 9.2 in aller Ausführlichkeit darlegt unter dem Titel der „Profilanalyse“ (S.275). In Kasten 21-4 werden alle drei Schritte exemplarisch durchgespielt.

Drei Beispiele zum Untersuchungsbericht

- Es folgen drei Beispiele:
- zu einem Leistungstest

in Kasten 21-4,
- zu einem Fragebogen

in Kasten 21-5,
- zu einem projektiven Verfahren

in Kasten 21-6.

Dazu der **HINWEIS**: Schritt 1 und 2 sind „handwerkliche“ Auswertungsleistungen, die den vollständigen Untersuchungsbericht vorbereiten. Die Ergebnisse aus Schritt 1 und 2 bleiben in den Unterlagen des Untersuchers. *Im Gutachten erscheinen nur die Resultate aus Schritt 3.*

Erstes Beispiel für einen Untersuchungsbericht
Es geht um einen Leistungstest:

Kasten 21-4:
Untersuchungsbericht zu einem Leistungstest

Die Daten stammen aus dem „Leistungs-Prüfsystem (LPS)“ von Horn (1983).
Der Proband, Herr Natter, ist 24 Jahre alt.

Schritt 1 eines Untersuchungsberichtes
Quantitative Auswertung

Das Verfahren wird ausgewertet nach Zahlen und Kürzeln, die einzelnen Test-Scores werden ermittelt. Diese Test-Scores werden nach zwei Gesichtspunkten klassifiziert:

- nach dem Mittelwert der Normgruppe

(A)
- nach dem individuellen Mittelwert des Probanden

(B)

(A)
Vergleich mit Normgruppen-Mittelwert

Ermittelt werden zunächst die Rohwerte des Probanden. Dann werden die Rohwerte umgewandelt in Normwerte: Beim LPS handelt es sich um „Centile“ (C) mit dem Mittelwert C = 5 und der Standardabweichung C = 2. Nachgeschlagen wird im Manual bei der entsprechenden Normgruppe, hier der Gruppe der 24jährigen.

In einer Übersicht werden die Ergebnisse zusammengestellt und die Test-Scores nach ihrer Ausprägung klassifiziert.

Als **Kürzel** werden verwendet:
UT : Untertest,
RW: Rohwert,
PR : Prozentrang,
C : Centilwert (M = 5 und SD = 2),
GL : Gesamtleistung im LPS,
VB : Vertrauensbereich ($p \leq 5\%$)
(Berechnung: S. 89 und 278!)

Die **Ausprägung** wird eingestuft nach drei Hauptklassen:
d : durchschnittlich
(C = 3 bis C = 7),
ü-d : überdurchschnittlich
(C > 7),
u-d : unterdurchschnittlich
(C < 3)

Doch beachte! Wenn Werte an der Grenze zweier Klassen liegen (beispielsweise bei $C = 3$), sei die Lage eigens erwähnt. Warum? Punktuell genommen gehört $C = 3$ zum Durchschnittsintervall. Beachtet man jedoch den Vertrauensbereich, dann kann der „wahre Wert“ für $C = 3$ im durchschnittlichen, aber auch im unterdurchschnittlichen Bereich liegen.

	UT		RW	C	VB	Ausprägung
1	+	2	57	6	6 ± 0.4	d (oberer Durchschnittsbereich)
		3	34	8		
		4	32	8		
3	+	4	66	8	8 ± 1.2	u-d (Grenze: ü-d / d)
		5	21	6		
		6	38	6		
5	+	6	65	6	6 ± 0.6	d (oberer Durchschnittsbereich)
		7	23	6		
		8	31	6		
		9	33	8		
		10	35	8		
7	bis	10	122	7	1 ± 0.4	d (Grenze: d / ü-d)
		11	27	6		
		12	26	5		
11	+	12	53	5	5 ± 0.8	d
		13	14	5		
		14	28	6		
13	+	14	42	6	6 ± 0.4	d (oberer Durchschnittsbereich)
	GL		405	7.6	7.6 ± 0.4	ü-d
	PR		90.3			
	IQ		118-121			

(B)

Vergleich mit dem individuellen Mittelwert des Probanden

Wir vergleichen die Ergebnisse der einzelnen Untertests mit dem individuellen Mittelwert des Probanden. Als „individuellen Mittelwert“ (IM) berechnen wir das arithmetische Mittel der sechs Untertests (C-Werte): $[(1 + 2) + \dots + (13 + 14)] / 6 = 6.33$.

Hinweis: Wir gehen hier also anders vor als in Kapitel 9.2 (S.278); dort hatten wir die Gesamtleistung (GL) als „individuellen Mittelwert“ (IM) gewählt. Warum? In Kapitel 9.2 liegen GL und IM dicht beieinander ($GL = 6$, $IM = 5.6$). Hier dagegen weichen GL und IM erheblich voneinander ab ($GL = 7.6$, $IM = 6.3$). Darum gehen wir hier von dem tatsächlichen „individuellen Mittelwert“ aus, von $IM = 6.3$.

Warum divergieren hier GL und IM erheblich? Bei der Transformation der Rohwerte in C-Werte bleibt eine Anzahl Rohwerte unbeachtet, sie gehen gleichsam „verloren“. - Beispiel: Im UT (1+2) hat Herr Natter 57 Rohpunkte erreicht. Er erhält dafür einen C-Wen von 6. Für einen C-Wert von 6 hätten 47 Rohpunkte gereicht. 10 Rohpunkte bleiben unberücksichtigt - sie sind gleichsam „überschüssig“. - Die „überschüssigen“ Rohpunkte gehen jedoch in die Summe der Rohpunkte bei GL ein. Diese Gesamtsumme der Rohpunkte ergibt darum für GL einen höheren C-Wen als die Mittelung aller C-Werte über die sechs Untertests.

Der Vergleich mit dem individuellen Mittelwert (IM) dient dazu, das sogenannte „individuelle Leistungsprofil“ zu ermitteln. Dazu bedarf es der Berechnung der Kritischen Differenzen Berechnung: S. 89 und 278!)

Es seien ermittelt:

für die Gesamtleistung (GL):

- der Standardmeßfehler: $SE_{GL} = 0.2$,
- der Vertrauensbereich: $VB_{GL} = 7.6 \pm 0.4$ ($p \pm 5\%$);

2. die Kritischen Differenzen zwischen Untertests und individuellen Mittelwert.

IM : Individueller Mittelwert UT : Untertest ED : Empirische Differenz ($C_{UT} - C_{IM}$) KD : Kritische Differenz					Hoch : Leistungshoch Tief : Leistungstief NS : Nicht signifikant (We- der Hoch noch Tief)		
					Interpretation		
UT	r_{tt}	C-Wert	ED	KD	Hoch	Tief	NS
(1 + 2)	0.96	6	-0.3	1.10			X
(3 + 4)	0.90	8	1.7	1.46	X		
(5 + 6)	0.98	6	-0.3	0.96			X
(7 bis 10)	0.99	7	1.0	0.87	X		
(11 + 12)	0.96	5	-1.3	1.10		X	
(13 + 14)	0.99	6	-0.3	0.87			X
IM	0.96*						

* Reliabilität für IM: Mittel über die Reliabilitätswerte der 6 UT.

Schritt 2 eines Untersuchungsberichtes

Stichwortartige Interpretation

HINWEIS: Die Ergebnisse aus Schritt 1 werden in Stichworten interpretiert, unter Beachtung des Vertrauensbereiches.

(1) Vergleich mit den Altersnormen

UT 1 + 2: **Allgemeinbildung, Wortverständnis:**
durchschnittlich (oberer Durchschnittsbereich)

UT 3+4: **Denkfähigkeit:**
überdurchschnittlich (Grenze überdurchschnittlich / durchschnittlich)

UT 5 + 6: **Wortflüssigkeit:**
durchschnittlich (oberer Durchschnittsbereich)

UT 7 bis 10: **Räumliche Vorstellung:**
durchschnittlich (Grenze durchschnittlich / überdurchschnittlich)

UT 11 + 12: **Gestaltbindung, Ratefähigkeit:**
durchschnittlich

UT 13 + 14: **Wahrnehmungstempo:**
durchschnittlich (oberer Durchschnittsbereich)

GL: **Gesamtintelligenz:** überdurchschnittlich

(2) Vergleich mit dem individuellen Mittelwert

Individuelles Leistungshoch:	Denkfähigkeit und Räumliche Vorstellung
Individuelles Leistungstief:	Gestaltbindung oder Ratefähigkeit

Schritt 3 eines Untersuchungsberichtes

Vollständiger Text

HINWEIS: Aus den Ergebnissen der Schritte 1 und 2 wird ein fortlaufender Text erstellt - nach den Vorschlägen, die in den vorhergehenden Abschnitten besprochen wurden.

Testbeschreibung: Das Leistungs-Prüf-System von Horn (LPS) soll wichtige Grundfähigkeiten der Intelligenz erfassen. Repräsentiert werden diese Dimensionen in unterschiedlichen Aufgabengruppen, die zu Untertests zusammengefaßt werden. Normen hegen vor für verschiedene Altersgruppen.

Verhaltensbeobachtung: Herr Natter verstand die Instruktionen sofort, keine mußte wiederholt werden...

Ergebnisteil: Herr Natter erzielte in den einzelnen Untertests (UT) folgende Werte, mitgeteilt in Centilwerten (C: Mittelwert = 5, Standardabweichung = 2). Die Vergleiche beziehen sich auf die Altersnormen. - Für die einzelnen C-Werte sei der *Vertrauensbereich* (VB) mitbestimmt und bei Angabe der Merkmalsausprägung mitberücksichtigt; die Restwahrscheinlichkeit beträgt fünf Prozent. - Die *Merkmalsausprägung* wird nur angegeben in den drei Klassen: durchschnittlich (d), unterdurchschnittlich (u-d) und überdurchschnittlich (ü-d). Wenn jedoch Werte an der Grenze zweier Klassen liegen (etwa bei C = 7) wird die Grenzlage miterwähnt.

<i>UT</i>	<i>Erfaßtes Merkmal</i>	<i>c</i>	<i>VB</i>	<i>Ausprägung</i>
1 + 2	Allgemeinbildung/Wortverständnis	6	6 ± 0.4	d
3 + 4	Denkfähigkeit	8	8 ± 1.2	ü-d (Grenze d)
5 + 6	Wortflüssigkeit	6	6 ± 0.6	d
7 bis 10	Räumliche Vorstellung	7	7 ± 0.4	d (Grenze ü-d)
11 + 12	Gestaltbindung/Ratefähigkeit	5	5 ± 0.8	d
13 + 14	Wahrnehmungstempo	6	6 + 0.4	d

Faßt man die Leistungen über alle Untertests zusammen und berechnet eine *Gesamtleistung*, so ergibt sich ein C-Wert von 7,6, das entspricht einem IQ von 118-121 und einem Prozentrang von 90,3. (Dieser Prozentrang besagt, daß nur noch rund neun Prozent der Altersgruppe höhere Leistungen erbringen wurden.)

Beachtet man bei der *Gesamtleistung* eine Irrtumswahrscheinlichkeit von fünf Prozent, so liegt der wahre Wert in dem Intervall von C = 7.2 bis C = 8.0. (Eine Irrtumswahrscheinlichkeit von fünf Prozent besagt: Wird der Test hundertmal angewandt, so ist höchstens in fünf Fällen damit zu rechnen, daß der wahre Wert außerhalb des berechneten Intervalles liegt.)

Interpretation: Die Gesamtintelligenz ließ sich als überdurchschnittlich einstufen. Der Vergleich bezieht sich auf die Altersnormen.

Im Vergleich zur Altersgruppe galt ferner:

- Im *überdurchschnittlichen* Bereich (an der Grenze zum Durchschnittsbereich) lag die Denkfähigkeit.
- Als *durchschnittlich bis überdurchschnittlich* erwies sich die Räumliche Vorstellung.
- *Durchschnittlich* waren
- die Allgemeinbildung (oberer Durchschnittsbereich),
- die Wortflüssigkeit (oberer Durchschnittsbereich),
- das Wahrnehmungstempo (oberer Durchschnittsbereich),
- die Gestaltbindung.

Das individuelle Leistungsprofil ergab bedeutsame Abweichungen zwischen der mittleren Gesamtleistung und einzelnen Teilfunktionen:

- Individuelle Leistungshochs fanden sich bei Denkfähigkeit und Räumlicher Vorstellung.
- Ein individuelles Leistungstief zeigte sich bei Gestaltbindung oder Ratefähigkeit.

Zweites Beispiel für einen Untersuchungsbericht
Es geht um einen Persönlichkeitstest:

Kasten 21-5:
Untersuchungsbericht zu einem Persönlichkeitstest

*Die Daten stammen aus dem „Freiburger Persönlichkeitsinventur (FPI)“
von Fahrenberg, Selg und Humpel (1978)
Der Proband, Herr Bunt, ist 28 Jahre alt.*

Testbeschreibung: Beim „Freiburger Persönlichkeitsinventar (FPI)“ handelt es sich um einen Fragebogen, der aus Antworten des Probanden zwölf Persönlichkeitsmerkmale ableitet. Normen liegen für Männer und Frauen sowie für unterschiedliche Altersgruppen vor.

Verhaltensbeobachtung: Herr Bunt beantwortete die Fragen sehr rasch, ohne jede Unterbrechung. Zu einzelnen Sätzen gab er, wie nebenbei, ironische Kommentare, versicherte aber spontan bei Abgabe des Antwortbogens: „Nehmen Sie meine Bemerkungen nicht zu ernst, ich habe so ehrlich geantwortet, wie ich konnte.“

Ergebnisteil: Herrn Bunt wurde die Langform des FPI vorgelegt. Es ergaben sich die folgenden Stanine-Werte (ST). (Der Mittelwert liegt bei ST = 5, die Standardabweichung beträgt ST = 2.) Der Vergleich bezieht sich auf die Altersgruppe. - Für die einzelnen Stanine-Werte sei der *Vertrauensbereich* (VB: nach Tabelle 26 des Manuals) mitangegeben und bei der Merkmalsausprägung mitberücksichtigt; die Restwahrscheinlichkeit beträgt fünf Prozent. - Die **Merkmalsausprägung** wird nur angegeben in den drei Klassen: durchschnittlich (d), unterdurchschnittlich (u-d) und überdurchschnittlich (ü-d). Wenn jedoch Werte an der Grenze zweier Klassen liegen (etwa bei ST = 7) wird die Grenzlage miterwähnt.*

Skala	Merkmal	ST	VB	Ausprägung
1	Nervosität	6	6 ± 1.3	d (Grenze: d / ü-d)
2	Aggressivität	7	7 ± 1.6	d (Grenze: d / ü-d)
3	Depressivität	5	5 ± 1.2	d
4	Erregbarkeit	6	6 ± 1.2	d (Grenze: d / ü-d)
5	Geselligkeit	9	9 ± 1.5	
6	Gelassenheit	8	8 ± 1.8	ü-d (Grenze: ü-d / d)
7	Dominanz	3	3 ± 1.8	d (Grenze: d / u-d)
8	Gehemmtheit	6	6 ± 1.5	d (Grenze: d / ü-d)
9	Offenheit	8	8 ± 1.8	ü-d (Grenze: ü-d / d)
E	Extraversion	9	9 ± 1.7	ü-d
N	Neurotizismus	4	4 ± 1.5	d (Grenze: d / u-d)
M	Maskulinität	1	1 ± 1.8	u-d

Zwischenbemerkung: Zur Interpretation

Wie ausführlich die Interpretation gefußt wird, soll der Untersucher von der Fragestellung her entscheiden. Wir gehen hier auf zwei Fälle ein. (Die praktische Arbeit, etwa in einer Erziehungsberatungsstelle, stellt den Diagnostiker vermutlich vor mehr als zwei solcher Fragestellungen!)

Fall A: In dem einen Falle genügt es, nur „auffällige“ Werte zu interpretieren. Was als „auffällig“ gelten soll, muß der Untersucher im Einzelfall von der Fragestellung her festlegen.

Fall B: In dem anderen Falle ist es dringend geboten, auch die „durchschnittlichen“ Merkmalsausprägungen aufzuführen. Beispielsweise kann es in einem forensischen Fall bedeutsam sein, darauf hinzuweisen, daß der Angeklagte „nur durchschnittliche“ Werte für Aggressivität erhalten hat.

Wir bieten eine Interpretation für Fall A und Fall B.

Interpretation: Fall A***Interpretiert werden nur „auffällige“ Werte***

HINWEIS: Als „auffällig“ bestimmen wir hier Merkmale, die über- oder unterdurchschnittlich ausgeprägt sind, und solche Merkmale, die - gemäß Vertrauensintervall - in den über- oder unterdurchschnittlichen Bereich hineinreichen können. Wir interpretieren also Werte, von denen gilt: $ST \geq 7$ oder $3 \leq ST$

Interpretation: Die Ergebnisse lassen die Bereitschaft erkennen, eigene (kleinere) Schwächen offen anzuerkennen. Das bedeutet: Die Fragebogenwerte sind verwert- und interpretierbar.

Nach diesen Ergebnissen war die Merkmalsausprägung

- überdurchschnittlich bei Geselligkeit und Extraversion,
- überdurchschnittlich bis durchschnittlich bei Dominanz,
- durchschnittlich bis überdurchschnittlich bei Aggressivität,
- durchschnittlich bis unterdurchschnittlich bei Dominanz,
- unterdurchschnittlich bei Maskulinität.

Interpretation: Fall B***Interpretiert werden alle Werte***

HINWEIS: Wir gliedern die Interpretation nach den drei Bereichen „überdurchschnittlich“, „durchschnittlich“, „unterdurchschnittlich“, Die Hinweise zur Grenzlage - gemäß Vertrauensbereich - fügen wir in Klammern hinzu, dabei verwenden wir um der Übersicht willen die drei Kürzel ü-d, d, u-d.

Interpretation: Die Ergebnisse lassen die Bereitschaft erkennen, eigene (kleinere) Schwächen offen anzuerkennen. Das heißt, die Testergebnisse sind interpretierbar.

Überdurchschnittlich ausgeprägt waren

- Geselligkeit,
- Extraversion,
- Gelassenheit (Grenze ü-d / d).

Als durchschnittlich erwiesen sich die Merkmale

- Aggressivität (Grenze d / ü-d),
- Nervosität (Grenze d / ü-d),
- Erregbarkeit (Grenze d / ü-d),
- Gehemmtheit (Grenze d / ü-d),
- Depressivität,
- Neurotizismus (Grenze d / u-d) und
- Dominanz (Grenze d / u-d).

Unterdurchschnittlich ausgeprägt war das Merkmal

- Maskulinität.

Drittes Beispiel für einen Untersuchungsbericht

Es geht es um ein projektives Verfahren.

Kasten 21-6:**Untersuchungsbericht zu einem projektiven Verfahren**

Die Angaben stammen aus dem „Thematischen Apperzeptionstest (TAT)“ von Murray (1943).

HINWEIS: Die ‚unpersönliche Prädikation‘ wird in der Weise gewahrt, daß nur von Inhalten und Figuren der Geschichten berichtet wird, nicht unmittelbar vom Erleben des erzählenden Probanden.

Testbeschreibung: Beim Thematischen Apperzeptionstest von Murray (TAT) soll der Proband zu verschiedenen mehrdeutigen Bildtafeln dramatische Geschichten erzählen. Eine Analyse der Inhalte kann Hinweise auf das Umwelterleben, auf Gefühle und Einstellungen des Probanden erbringen.

Verhaltensbeobachtung: Herr X. sprach sehr leise, machte viele Pausen und betonte immer wieder, wie schwer es ihm falle, Geschichten zu erfinden, weil er ein phantasieloser Mensch sei...

Ergebnisbericht: entfällt - wie meist bei projektiven Verfahren.

Interpretation: Aus den erzählten Geschichten ließen sich Hinweise zu folgenden Inhalten ableiten:

- Eher gedruckte Stimmungslage,
- Aggressive Impulse gegenüber nahestehenden Personen,
- Unsicherheit, wenn es gilt, neue Kontakte aufzunehmen,
- Bestreben, für unakzeptierte Wünsche Scheingründe vorzuschützen (Rationalisierungstendenz),
- Auseinandersetzung mit den Themen:
 - ⇒ „Isoliert zu sein, auf sich allein gestellt zu sein“,
 - ⇒ „Konflikte aus dem Wege zu gehen“,
 - ⇒ „Immer wieder zu erleben, daß Partnerschaften zerbrechen“,
 - ⇒ „Sich selbst behaupten zu wollen gegenüber Autoritäten“.

Kommentar zu Kasten 21-6: Aussagen, wie hier präsentiert in der Interpretation, erhalten ihren Sinn - sozusagen „ihren Sitz im Leben“ - erst im Rahmen eines konkreten Falles.

21.4.4 Vierter Abschnitt des Gutachtens: **Befund**

Der Befund nimmt eine Schlüsselstellung im Gutachten ein: Was Vorgeschichte und Untersuchungsbericht in den einzelnen Verfahren referieren, soll der Befund thematisch zusammenfassen.

Aussagen zu demselben Thema, die in *verschiedenen* Verfahren stehen, werden im Befund in *einen* Abschnitt zusammengezogen.

Beispiel: Eine Exploration enthält Angaben über Sozialkontakte. Aussagen über Sozialverhalten können sich auch an anderer Stelle finden, etwa in Fragebogen oder in projektiven Verfahren. Alle Informationen über Sozialverhalten werden an einer „Stelle“ zusammengefaßt.

Wenn der Befund die Angaben zu denselben Merkmalen jeweils in einem Abschnitt zusammenzieht, dann dient er der *Integration* der Angaben aus Vorgeschichte oder Untersuchungsberichten, in diesem Sinne auch einer Reduktion der Redundanz. Aber er hat noch weitere Funktionen.

Die Untersuchungsberichte stellen das Verhalten des Probanden test- und situationsbezogen dar; dabei bleibt unentschieden, ob die erfaßten Merkmale zu dem (relativ konstanten) Verhaltensrepertoire des Probanden gehören. Im Befund soll der Gutachter weitergehen: Er soll Verhaltensanteile identifizieren,

von denen er annimmt, daß sie eine *relative Stabilität* besitzen (über eine gewisse Zeitspanne hin und über relevante Situationen hinweg). Soweit Situationsabhängigkeit erkennbar wird, soll sie im Befund miterwähnt werden.

Nach unserem Konzept kennzeichnet den Befund noch eine weitere Eigenart: Er bleibt *deskriptiv*, in ihm werden keine diagnostischen oder prognostischen Schlüsse gezogen.

Schließlich gilt als Regel für die Darstellung: Im Befund werden alle Aussagen *aus der Perspektive des Probanden* formuliert.

Somit charakterisieren den Befund vier Titel:

1. Integration von Aussagen,
2. Beschreibung von relativ stabilem Verhalten,
3. Deskription von Verhalten (nicht Explikation),
4. Darstellung aus der Perspektive des Probanden.

Zu 1.: Integration von Aussagen:

In derselben Untersuchung werden in der Regel unterschiedliche Verfahren verwandt, etwa Exploration/Anamnese einerseits und Leistungstests/Fragebogen andererseits. Lassen sich Aussagen aus so unterschiedlichen Verfahren zu *einer* Gesamtaussage integrieren? Als Antwort vier Hinweise:

- a) Der Gutachter ist es, dem eine zentrale Rolle für die Lösung des Integrationsproblems zufällt. Von ihm wird angenommen, er könne die Einzelverfahren verstehen, handhaben, auslegen und - *gewichten*: Aussagen auszuwählen heißt immer auch „Aussagen zu gewichten“.
- b) Die quantitativen Verfahren (etwa Tests oder Fragebögen) liefern über ihre Aussagemöglichkeiten und Aussagegrenzen eine Fülle von Informationen, wenn man sich die Mühe macht, die Handanweisungen zu studieren.
Bei Test und Fragebogen dürften für die integrative Aufgabe drei Klassen von Angaben informativ sein:
 - die theoretische Zuordnung eines Verfahrens,
 - die Höhe der konvergenten und divergenten Trennschärfen,
 - schließlich die Indizes der konvergenten und diskriminanten Validität (also der Zusammenhang zu Kriterien, die ähnliche oder abweichende Konzepte repräsentieren).
- c) Was die Informationen aus qualitativen Verfahren angeht (etwa aus einer Exploration oder aus dem TAT), so stellt sich die Frage der Integration nur für den, der ihren Einsatz für gerechtfertigt hält.
Zu einer Rechtfertigung zwei Überlegungen:
 - Der diagnostische Urteilsprozeß - auch, wenn er hochformalisiert ist
 - erfordert ein Mindestmaß an ‚offenem‘ Informationsaustausch zwischen dem Probanden und dem Gutachter: Der Proband nennt sein Problem, der Psychologe „übersetzt“ es in ein Untersuchungsdesign, er ordnet einzelnen Problemsegmenten bestimmte Verfahren zu und legt

für ihren Einsatz eine zeitliche Sequenz fest, schließlich transformiert er seine Ergebnisse zurück in die Sprache eines Laien usw. (siehe Kap. 19, S.413).

Die Urteilsschritte, die dabei zu tun sind, ähneln mehr dem Abwägungsprozeß, den qualitative Verfahren erfordern, als dem Vorgehen, das für psychometrische Instrumente vorgesehen ist.

- Wer qualitative Verfahren verwendet, muß nicht blind sein gegenüber ihren Schwächen und Nachteilen. Im Gegenteil, der „informierte“ Diagnostiker kann aufgrund seiner Kenntnisse (vermutlich) Chancen und Grenzen klarer abschätzen als der Psychologe, der ihre Verwendung ablehnt.

Qualitative Verfahren können bedeutsame Hinweise geben für die Generierung von Hypothesen, für die Einbettung und die Interpretation von Ergebnissen aus psychometrischen Verfahren, für einen zusätzlichen Einsatz von Leistungs- oder Persönlichkeitstests.

- d) Hilfreich für die Integration verschiedener Aussagen ist eine Orientierung an den Elementarkategorien: bei einer Exploration etwa an *verhaltensnahen Angaben*, bei einem Test an den *Item-Inhalten*.

Beispiel: *Interessen lassen sich erfassen in einer Exploration, aber auch mit einem Fragebogen, etwa dem DIT¹. Nun werde bei einem Probanden in beiden Verfahren ein hohes „Interesse für Literatur“ erkennbar. Ob die Aussagen der Exploration und die Antworten zum DIT zueinander passen, kann der Gutachter nur entscheiden, wenn er die „Elementarkategorien“ vergleicht: die Einzelaussagen der Exploration und die Item-Inhalte des Fragebogens. - Nun berichte der Proband im Gespräch, er bevorzuge Schriftsteller wie Agatha Christie, Konsalik und Simmel. - Dann drückt sich darin ein **anderes** „Interesse für Literatur“ aus, als es die DIT-Items umschreiben; die DIT-Items charakterisieren „Interesse für Literatur“ durch Titel wie Sprachkritik, Beschäftigung mit gehobener Literatur Studium der Literaturgeschichte.*

Die vier ‚Hinweise‘ lassen sich in dem folgenden Satz zusammenfassen: Ein Gutachter, der sich vertraut macht mit Einzelverfahren, dürfte in der Lage sein, Einzelaussagen zu integrieren,

- wenn er anhand der Kennwerte die Aussagemöglichkeiten psychometrischer Instrumente abschätzt,
- dabei jedoch auf die Übersetzungs- und Verbindungsfunktion qualitativer Verfahren nicht verzichtet und
- gegebenenfalls die Elementarkategorien seiner Informationsquellen zur Aufhellung der Aussagen heranzieht.

Hinweis: *Einen solchen Urteilsprozeß für möglich zu halten muß nicht dazu verführen, seine Schwierigkeit zu unterschätzen. Auch nach sorgfältiger Abwägung lassen sich in den Aussagen die aufgehellten und die*

¹ DIT: „Differentieller Interessentest“ von Todt (1967)

‚dunklen‘ oder ‚leeren‘ Stellen nicht disjunkt trennen. Der Gutachter kann das Aussagenrisiko nur reduzieren, nicht eliminieren.

Zu 2.: Beschreibung von relativ stabilen Verhaltensanteilen:

Wer prinzipiell die Frage der Aussagen-Integration für lösbar hält, muß sich mit dem zweiten Hauptproblem einer Befunderstellung befassen: der Bestimmung der relativen Stabilität des Verhaltens.

Auf welche Weise läßt sich eine solche Stabilität bestimmen, daß sich Aussagen für den Befund ergeben? In Anlehnung an Thomae (1967, 747) seien drei Kriterien genannt:

- Erstens, *dasselbe Merkmal wird an verschiedenen Punkten der Zeitachse identifiziert.* Zu verschiedenen Zeitpunkten muß also dieselbe Situationsklasse erkennbar und das gleiche Verhalten beobachtbar sein. - Informationsquellen können in diesem Sinne sein: Vorgeschichte, explorative Methoden, biographische Inventare.
- Zweitens, *dasselbe Merkmal wird simultan in ähnlichen Situationsklassen festgestellt.* „Simultan“ bezieht sich dabei auf Ereignisse, die auf der Zeitachse in einem so kurzen Intervall plaziert sind, daß sie als „gleichzeitig“ interpretiert werden. - **Beispiel:** Dasselbe Merkmal, etwa Konzentration, wird in derselben Untersuchungsphase mit verschiedenen Verfahren gemessen.

Doppelbelege: *Von dieser Überlegung her rechtfertigt sich der Vorschlag, in der Regel jede Befundaussage durch Ergebnisse aus wenigstens zwei Verfahren (durch zwei „Belege“) zu stützen. - Die negative Variante dieses Argumentes würde lauten: Jedes psychologische Verfahren ist fehlerbehaftet. Wenn jedoch zwei unterschiedliche Verfahren **unabhängig voneinander** das gleiche Merkmal anzeigen, dann wächst die Wahrscheinlichkeit, daß ein systematischer nicht bloß ein zufälliger Meßeffect vorliegt.*

Es gibt Ausnahmen: *Scores aus bewährten psychometrischen Verfahren oder spezielle „unerfindbare“ Details aus Gesprächen können auch als Einzelbelege auftreten!*

Gewichtung: *Bevor der Gutachter zwei „Belege“ (also Ergebnisse zweier Verfahren) in einer Aussage zusammenfaßt, sollte er ihr „diagnostisches Gewicht“ einstufen:*

- ⇒ *Als „stark“ dürften Belege gelten, die aus bewährten psychometrischen Verfahren stammen. - Dabei sind Fragen zu klären wie die folgenden: Welche Kennwerte liegen vor? Gibt es Normen, die speziell auf den Probanden und seine Fragestellung passen?*
- ⇒ *Als eher ‚schwach‘ dürften Belege gelten, die aus Gesprächen oder aus projektiven Verfahren (etwa dem Baumtest²) stammen.*

2 Beim Baum-Test von Koch (1972) soll der Proband einen Baum zeichnen. Zur Auswertung legt Koch eigene Analysen von Zeichnungen vor und gibt Hinweise zur Deutung einzelner Merkmale.

- Drittens, das Merkmal, um das es geht, wird anderen Merkmalen zugeordnet, zu ähnlichen Eigenschaften müssen konvergente Beziehungen, zu unähnlichen dagegen diskriminante erkennbar werden. **-Beispiel:** *Ein Proband erhalte hohe Werte in „Depressivität“. Im Sinne der Konvergenz sollten dann ähnliche Werte auftauchen bei Merkmalen wie Hoffnungslosigkeit, Rigidität, Externale Kontrollüberzeugung³ usw. -Im Sinne der Diskordanz sollten niedrige Werte erscheinen bei Merkmalen wie Aggressivität, Gelassenheit, Geselligkeit usw.*

Bei allen drei Kriterien ist der Gutachter als urteilendes Subjekt die entscheidende Größe. Enden wir damit bei einem puren Subjektivismus? Ich meine: Nein! Beschränkt auf den subjektiven Geltungsbereich bliebe der Befund nur dann, wenn der Gutachter über seine Erkenntnisschritte keine Rechenschaft ablegen könnte, die intersubjektiv nachvollziehbar ist. **Eine intersubjektiv nachvollziehbare Rechenschaft wird aber für jeden Gutachtenschritt gefordert.**

Zu 3.: Deskription:

Warum soll der Befund rein deskriptiv bleiben? Der Befund zieht eine Summe aus den Informationen, die in Vorgeschichte und Untersuchungsberichten vorliegen. Ein solches Resümee zu bilden ist ein sehr komplexer Schritt. Man sollte ihn nicht weiter komplizieren, indem man zusätzliche Aufgaben hinzunimmt: z.B. eine diagnostische oder prognostische Aufhellung der Aussagen hinzufügt. Vielmehr empfiehlt es sich, diese ‚Erklärung‘ vom Befund zu trennen: der nächste Abschnitt, die Stellungnahme, ist dafür vorgesehen.

Zu 4.: Darstellung aus der Perspektive des Probanden:

Im Befund werden alle Angaben so geordnet, daß sie *den Probanden charakterisieren*. Dabei werden alle Informationen gleichsam ständig mit Blick auf den Probanden formuliert - nicht neutral wie in einer reinen Sachbeschreibung.

Zwei Beispiele:

- a) *Aus der Biographie von Frau Gründel sei bekannt, daß Partnerschafterz, die sie einging, immer wieder zerbrochen sind. Dieser Sachverhalt läßt sich völlig neutral referieren: „Partner verließen sie immer wieder“ Den Gutachter interessiert indessen, wie dieser Vorgang im Erleben von Frau Gründel aussieht. Er wird darum formulieren: „Frau Gründel befaßt sich intensiv mit der Frage, warum Partnerschaften, die sie begonnen hatte, immer wieder zerbrachen.“*
- b) *Die Umwelt erlebe Frau Gründel als kontaktscheu, **Falsche Formulierung** (nach unserer Konzeption): „Die Bekannten und Freunde beschreiben Frau Gründel als wenig spontan, als stark kontrolliert und distanziert im Kontakt.“ **Richtig - wenn belegbar:** „Frau Gründel verhält sich wenig*

³ „Externale Kontrollüberzeugung“ betrifft die Annahme, daß ein Individuum sein Verhalten stärker bestimmt sieht von Faktoren außerhalb als innerhalb der eigenen Person.

spontan, im Kontakt kontrolliert sie sich in hohem Maße und zeigt sich eher gehemmt.“

Die vier Charakteristika (Integration, Stabilität, Deskription, Probandenperspektive) legen es nahe, für die Abfassung des Befundes einige Sprachregeln zu formulieren (Kasten 21-7).

Kasten 21-7:
Abfassungsregeln zum Befund

- Zu dem gleichen Thema werden Aussagen aus verschiedenen Verfahren an *einer* Stelle zusammengezogen.
- Welche Themengruppen vorkommen, ist festgelegt mit der *Fragestellung* und mit der Auswahl des *Instrumentars*.
- Wie die Themengruppen im Befund *ungeordnet* werden, liegt im Ermessen des Gutachters.
- Eine *Gliederung* könnte sich orientieren an Konzepten wie:
 - ⇒ Intelligenz, gegebenenfalls gegliedert nach Teilfunktionen,
 - ⇒ Andere Fähigkeiten und Fertigkeiten,
 - ⇒ Interessen, Freizeittätigkeiten,
 - ⇒ Soziale Beziehungen
 - ⇒ usw.
- Es erscheinen *nur Beschreibungsdimensionen, die genereller Art sind* (Dispositionsprädikate), keine Einzelbeobachtungen mehr. (Es heißt nicht, Peter habe in der letzten Mathematikarbeit eine Fünf gehabt, sondern [falls belegbar], seine Leistungen in Mathematik seien unterdurchschnittlich.)
- Genannt werden vor allem auch *übergreifende Themen* (Thomae würde sie „Daseinsthemen“,⁴ nennen), etwa „Streben nach sozial gehobener Position“, „Bemühen um Daseinssteigerung, Kreativität, Selbstverwirklichung“ usw.
- Die Aussagen sind *nur feststellender, nicht erklärender Natur*. (Es wird nur gesagt, **daß** Peter sich gegenüber seinen Mitschülern aggressiv verhält, nicht **warum** er aggressiv ist.)
- Die *Diktion bleibt neutral*. Wertende Aussagen (wie etwa „gut“, „bedauerlich“, „verwerflich“) sind fehl am Platz, sie enthalten Stellungnahmen (oft in Gestalt impliziter Vorurteile).
- Weil die Aussagen Verhaltensweisen betreffen, die sich immer wieder bei der Person des Probanden beobachten lassen, ist die *Prädikation jetzt persönlich*. (Es heißt: „*Herr X ist überdurchschnittlich intelligent*.“ Es heißt nicht: Die Ergebnisse sprechen für eine überdurchschnittliche Intelligenz.)
- Das Tempus ist in der Regel das *Präsens*.
- In der Regel wird *eine Befindaussage durch zwei Belege gestützt*. In wenigen Fällen, bei psychometrisch vorzüglich bewährten Verfahren, genügt *ein* Beleg.
- Die Belege werden bei Erstellung des Befundes - am Textrand - aufgeführt, damit der Gutachter seine eigenen Aussagen kontrollieren und ein ‚Kollege‘ sie nachvollziehen kann.

⁴ „Daseinsthemen“ bezeichnen Leit motive des Verhaltens, sie lassen sich verstehen als motivational-kognitive Orientierungssysteme, aus denen heraus Individuen ihre Sinnsuche steuern (Thomae, 1988, 53).

Drei Erstellungsschritte

Da beim Befund eine Vielzahl von Informationen erscheint, empfiehlt es sich, sie in drei Schritten zu ordnen:

1. In einer **Befundliste** lassen sich die Einzelergebnisse stichwortartig zu thematischen Einheiten gruppieren.
2. In einer **Befundskizze** werden die Aussagengruppen der Befundliste reduziert.
3. Aus der Befundskizze wird ein fortlaufender **Text des Befundes** formuliert.

Diese drei Schritte seien zuerst skizziert, dann an Beispielen veranschaulicht.

Erster Schritt einer Befunderstellung:

Sammlung von Daten für eine Befundliste

In der Vorgeschichte und in den Untersuchungsberichten sind die Informationen nach Verfahren gruppiert. Die Befundliste dient dem Ziel, die *Informationen* in einem ersten Schritt *nach Themen* zu ordnen. Im Befund sollen die Informationen nach Themen zusammengefaßt werden.

Demnach muß der Gutachter die Informationen der Vorgeschichte und der Untersuchungsberichte in größere Themenbereiche aufgliedern. Die Fragestellung und die Verfahrensliste der ‚Übersicht‘ (des ersten Gutachtenabschnittes) müßten ihm Stichworte vorgeben.

Hinweis: *Um die Übersicht zu erleichtern, kann er Aussagen zu gleichen Themen je auf einen Papierbogen übertragen oder je in eine edv-Datei einspeisen.*

Der Aufwand, den die Erstellung einer Befundliste erzwingt, rechtfertigt sich allein aus dem Grad der Kontrolle, den sie ermöglicht. Wer diese Kontrolle auf anderen Wegen ausübt, kann auf eine Befundliste verzichten.

Zweiter Schritt einer Befunderstellung:

Entwurf einer Befundskizze

In der Befundliste können Daten nebeneinanderstehen, die aus unterschiedlichen Verfahren stammen, aber dasselbe Merkmal betreffen. In der Befundskizze integriert der Gutachter die Informationen aus der Befundliste. Beispielsweise können in der Befundliste zu dem Merkmal ‚Intelligenz‘ Angaben aus vier Verfahren zusammentreffen; in der Befundskizze werden die unterschiedlichen Daten zur „Intelligenz“ in einem Abschnitt integriert.

Was „Integration“ bedeutet, wurde in den Anfangsabschnitten dieses Teilkapitels erläutert. Wiederholt sei hier das Stichwort, daß Integration immer Auswahl und „Gewichtung“ einschließt. Beispiel für die Gewichtung: Enthält ein Leistungstest Angaben zur Höhe der Intelligenz, so kommt diesem Test-Score

ein „höheres Gewicht“ zu als der Angabe aus dem Rorschach oder aus einem Selbstbericht.

Dritter Schritt einer Befunderstellung:

Befundformulierung

Befundliste und Befundskizze haben nur Dienstfunktion, sie sollen die Formulierung des Befundes erleichtern. Nur *der Befund erscheint im Originalgutachten*.

Die Abfassungsregeln wurden in Kasten 21-7 zusammengestellt (S. 461).

Es folgt ein Beispiel, das alle drei Schritte der Befunderstellung veranschaulicht.

Beispiel zur Befunderstellung

Fragestellung: Fahrtauglichkeit

Das Beispiel bezieht sich auf Herrn V., der den Führerschein „wegen Alkohols am Steuer“ verloren hat und einen Antrag auf Wiedererteilung der Fahrerlaubnis gestellt hat. (vgl. Beispiel 2 zur Vorgeschichte, S.446.)⁵.

(1) Befundliste

Die Befundliste führt die Aussagen aus Vorgeschichte und Untersuchungsbericht stichwortartig nach Themenkreisen zusammen.

Die Verfahren mit ihren Kürzeln werden in der Reihenfolge, in der sie in der Befundliste erscheinen, aufgeführt. Weniger bekannte Verfahren werden kurz erläutert:

HAWIE: Hamburg-Wechsler-Intelligenztest für Erwachsene
(GT: Gesamttest; VT: Verbalteil; HT: Handlungsteil. Vgl. Wechsler, 1964.)

BT: Benton-Test
(Figurale Darstellungen werden kurz dargeboten, sie sollen aus der Erinnerung nachgezeichnet werden. Geprüft werden Wahrnehmungs- und Verarbeitungsfunktionen. Vgl. Benton, 1972.)

TAVT: Verkehrsgebundener tachistoskopischer Auffassungsversuch
(Dem Probanden werden tachistoskopisch Verkehrssituationen dargeboten. Er soll verkehrsbedeutsame Einzelheiten erfassen.)

SOP: Sortierprobe nach Stein
(Durch die Deckelöffnungen eines Kastens sollen Körper geschoben werden, die den Öffnungen entsprechen. Geprüft wird Problemlöseverhalten bei einfachen Aufgaben.)

⁵ Zur Erinnerung: Es handelt sich um einen alten „echten“ Fall. Damm wurden die „alten“ Verfahren in der Aufzählung belassen.

- MU:** Visualitätsprobe nach Munsch (Verkehrs-Matrizentest)
(Abstrakte Muster werden dargeboten, die jedoch Verkehrsschildern und Verkehrswegen ähneln. In den Mustern ist ein Teil ausgespart, wie bei Matrizentests. Aus einer Reihe von ‚Ergänzungen‘ ist der richtige Teil auszusuchen.)
- KD:** Kieler Determinationsgerät nach Mierke
(Der Proband muß auf unterschiedliche optische und akustische Signale unterschiedlich reagieren. Geprüft wird die psycho-physische Belastbarkeit.)
- ZH:** Zweihand-Koordinationsprüfer
(Der Proband soll eine ‚Straße‘ nachfahren, die in eine Metallschablone gestanzt ist. Geprüft wird die visu-motorische Koordination.)
- VVT:** Verkehrs-Verständnis-Test von Müller
(Es werden Wissens- und Verständnisfragen über soziale und technische Aspekte der Verkehrssicherheit gestellt. Vgl. Müller, A., 1973.)
- E:** Exploration
- FPI:** Freiburger Persönlichkeitsinventar
(Vgl. Fahrenberg et al., 1978 und 1984.)
- TAT:** Thematischer Apperzeptionstest
(Vgl. Murray, 1943.)

1. Blatt: Intelligenz

Nr.	Aussage	Quelle
1	Gesamtintelligenz: durchschnittlich	HAWIE: GT
2	Mehr sprachlich gebundene Teilfunktion der Intelligenz: durchschnittl.	
3	Weniger sprachlich gebundene Teilfunktion der Intelligenz: durchschnittlich	HAWIE: VT HAWIE: HT
4	usw.	

2. Blatt: Andere Leistungsfunktionen

Nr.	Aussage	Quelle
8	Kein Hinweis auf Störungen im visuellen Wahrnehmungs- oder Verarbeitungs-bereich	BT
9	Schnelligkeit und Genauigkeit der verkehrsbezogenen visuellen Orientierung und Wahrnehmung: durchschnittlich (voll ausreichend)	TAVT
10	Visuelle Wahrnehmung von Details: voll ausreichend	SOP
11	Koordination zwischen visueller Wahrnehmung und manuell-motorischer Manipulation: voll ausreichend	SOP
12	Visualität im Sinne einer exakten Erfassung und Zuordnung visueller Details (verkehrsorientiert): voll ausreichend	MU
13	Sensu-motorische Mehrfachkoordinationsleistung: voll ausreichend	KD
14	Sensu-motorische Koordinationsleistung zwischen visueller Wahrnehmung und manuell-motorischer Steuerung: voll ausreichend	ZH
15	usw.	

5.Blatt: Verkehrsbezogene Einstellungen und Sichtweisen

Nr.	Aussage	Quelle
25	Problemverständnis für verkehrsbezogene Sachverhalte: überdurchschnittlich	VVT
26	Differenziertes Verständnis für Verkehrsprobleme	E
27	Bereitschaft, eigene Schuldanteile an Verkehrsvergehen oder Verkehrsunfällen zuzugeben	E
28	Tendenz, den Führerscheinentzug als ungerecht einzustufen	E
29	Eingeständnis von schuldhaftem Verhalten in Verkehrssituationen	TAT
30	Offenheitstendenz; Bereitschaft, eigene (kleine) Schwächen zuzugeben: überdurchschnittlich	FPI (9)
31	usw.	

(2) Befundskizze

In der Befundskizze werden aus der Befundliste Aussagen zu demselben Verhaltensbereich zusammengeführt.

Zur Befundskizze seien zwei Hinweise wiederholt:

1. In der Regel wird *eine Befundaussage durch zwei Belege gestützt*. In wenigen Fällen, bei psychometrisch vorzüglich bewährten Verfahren, genügt *ein Beleg*.
2. Integration schließt immer Auswahl und „Gewichtung“ ein.

Nr.	Aussage	Beleg/Quelle	Nr. in Befundliste
	Gesamtintelligenz: durchschnittlich	HAWIE: GT	
	Stärker auf sprachliche Inhalte bezogene Teilfunktion der Intelligenz: durchschnittlich	HAWIE: VT	2
	Weniger auf sprachliche Inhalte bezogene Teilfunktion der Intelligenz: durchschnittlich	HAWIE: HT	3
	Visualität im Sinne einer exakten Erfassung und Zuordnung optisch wahrgenommener Details: voll ausreichend	MU SOP TAVT	12 10 9
	Koordination zw. visueller Wahrnehmung und manuell-motorischer Steuerung: voll ausreichend	ZH SOP KD	14 11 13
	Differenziertes Verständnis für Verkehrsprobleme	VVT E	25 26
	Bereitschaft, eigene ‚Schuldanteile‘ an Fehlverhalten zuzugeben	FPI (9) E	30 27
	usw.		

(3) Ausformulierter Befund

Im Befund wird aus der Befundskizze ein fortlaufender Text erstellt - *dieser ausformulierte Text geht allein in das Gutachten ein*. Befundliste und Befundskizze bleiben in den Unterlagen des Gutachters.

Die Sprachregeln gibt Kasten 21-7 vor (S. 461).

Kasten 21-8:
Befundformulierung
Fragestellung: Rückgabe eines Führerscheins

Text	Quelle
Die Resultate der psychologischen Exploration, der Leistungs- und Persönlichkeitsuntersuchungen ergaben folgendes Bild:	
Der 37jährige Herr V. verfügt über eine durchschnittliche Gesamtintelligenz. Ebenfalls durchschnittlich ausgeprägt sind zwei intellektuelle Teilfunktionen: die stärker auf sprachliche Inhalte bezogenen Fähigkeiten und die weniger sprachorientierten Intelligenzbereiche.	HAWIE: GT HAWIE: VT HAWIE: HT
...	
Was die visuelle Orientierung angeht, so ist Herr V. in voll ausreichendem Maße in der Lage, Gegebenheiten und Beziehungen, auch in Details, visuell sicher zu erfassen.	BT, TAVT, SOP, MU
Auf diese visuellen Wahrnehmungen kann er seine manuell-motorische Reaktionen rasch und genau abstimmen.	SOP, KD, ZH
Für verkehrsbezogene Sachverhalte bekundet er ein differenziertes Verständnis, das die Problemsicht des Durchschnitts der Kraftfahrer übertrifft.	VVT E
Er zeigt sich bereit, auch Fehlverhalten zuzugeben, speziell im Straßenverkehr.	FPI(9) E, TAT

21.4.5 Fünfter Abschnitt des Gutachtens: Stellungnahme

Die Stellungnahme bietet die Antwort auf die diagnostische Frage. Herleiten muß sie sich aus den Informationen, die der Gutachter in Vorgeschichte, Untersuchungsbericht und Befund aufbereitet hat.

Die „Richtlinien für die Erstellung Psychologischer Gutachten“ (dpv, 1994 b, 11) skizzieren die Aufgabe wie folgt: „Die Stellungnahme soll die Problemlage sowie die Bedingungen für Entstehung und Aufrechterhaltung des Problems kenntlich machen. In vielen Fällen gehört es zum Gutachterauftrag, über diagnostische Feststellungen hinaus konkrete Maßnahmen vorzuschlagen; diese müssen schlüssig an die diagnostischen Befunde anknüpfen und dem aktuellen Stand der Forschung entsprechen.“

Hier **Abfassungsregeln** zu formulieren ist schwieriger als bei den anderen Gutachtenabschnitten. Orientieren muß sich die Stellungnahme vor allem - wie jeder andere Gutachtenteil - an der diagnostischen Frage:

- Um die Stellungnahme als Antwort zu kennzeichnen, kann es hilfreich sein, die diagnostische Frage wörtlich oder sinngemäß zu wiederholen.
- Für die Gliederung empfiehlt es sich, die Gesamtfrage (sofern möglich) in **Teilfragen** zu zerlegen und jede einzeln zu beantworten.
- Der Anfänger könnte sich die Antwort auch erleichtern, wenn er die Stellungnahme nach drei **Orientierungsfragen** gliedert:
 1. Welche Probleme liegen vor?
Es werden zentrale Themen oder Konflikte aufgelistet.
 2. Worauf gehen die Probleme zurück?
Es wird das Bedingungsgefüge gesucht, aus dem sich Themen oder Konflikte erklären lassen.
 3. Was kann geschehen, um die Probleme zu lösen?
Es werden Vorschläge (Interventionsvorschläge) formuliert, die dazu beitragen sollen, (die) Konflikte zu lösen. Welche Prognosen lassen sich mit den Vorschlägen verknüpfen?
- Wenn die „dritte Orientierungsfrage“ Interventionsvorschläge einschließt, dann bleibt abzuklären,
 - ⇒ erstens, welche *Ziele* anzustreben sind (Was kann oder soll geschehen?),
 - ⇒ zweitens, welche Methoden den Zielen affin sind? (*Wie* kann man die Ziele verwirklichen?)
- Um seine Argumentationskette sichtbar zu machen, sollte der Gutachter kontrollieren, ob sich Ursachenzuweisung, Vorhersage oder Entscheidungsvorschlag aus Vorgeschichte, Untersuchungsbericht und Befund begründen lassen. (Gegebenenfalls sollte er die Belegstellen an den Rand seines Textes setzen.)
- Der Gutachter sollte seine Antwort auf die Fragestellung beschränken (es sei denn, Gründe tauchen auf, die eine Ausweitung rechtfertigen).
- Um die Stellungnahme auch sprachlich als Antwort zu kennzeichnen, könnte er die entscheidenden Sätze in Entsprechung zur Ausgangsfrage formulieren.
- Wenn er zu keinem eindeutigen Urteil kommt, sollte er dies zum Ausdruck bringen.

Hinweis zur Intervention: Von allen Teilen des Gutachtens nähert sich die Stellungnahme am engsten den Absichten einer **Intervention**. Die Beispiele, die wir bringen, führen aus ihrem Anliegen heraus zu interventiven Handlungsvorschlägen.

Es folgen zwei **Beispiele**:

- das erste gliedert sich nach **Teilfragen**,
- das zweite nach drei **Orientierungsfragen**.

Erstes Beispiel zur Stellungnahme: *Gliederung nach Teilfragen*

Es geht um die Beurteilung der Fahrtauglichkeit von Herrn V., dem der Führerschein entzogen wurde „wegen fahrlässiger Straßenverkehrsgefährdung unter Alkoholeinfluß“ (vgl. Vorgeschichte und Befund, S. 463 und S. 446).

Der Gedankengang sei nur skizziert. Dabei tauchen auch Informationen auf, die **hier** in Vorgeschichte und Befund nicht aufgeführt waren (Kasten 21-9).

Kasten 21-9: Stellungnahme

Fragestellung: Wiedererteilung des Führerscheins

Gliederung: Der Gutachter wiederholt die Ausgangsfrage und zerlegt sie in zwei Teilfragen: Aussagen zur ‚Leistungsfähigkeit‘ und Aussagen zu ‚verkehrsrelevanten Einstellungen‘.

In einer weiteren Passage bespricht er die Alkoholproblematik.

Zum Abschluß erörtert er eine ‚Empfehlung‘ an die zuständige Behörde als Entsprechung zur Ausgangsfrage.

Hinweis: Wir fügen die Gliederungspunkte in Klammern zum Text hinzu. Im Original erscheinen sie nicht.

Text: Zur Frage steht, ob es geraten erscheint, Herrn V. die Fahrerlaubnis zurückzugeben. Das Problem sei unter zwei Perspektiven betrachtet: unter der seiner Leistungsfähigkeit und der seiner verkehrsrelevanten Einstellungen.

(Teilfrage I) Was das **Leistungsverhalten** betrifft, so erweisen sich die Fähigkeiten, die zum Kraftfahren erforderlich sind, in ihrer Ausprägung und in ihrem Übungsstand ausreichend ausgebildet. Intelligenz, Visualität und Konzentration sind wenigstens durchschnittlich ausgeprägt. Das gleiche gilt von Reaktionsschnelligkeit und psychophysischer Belastbarkeit. Damit verfügt Herr V. in diesen Funktionsbereichen über ausreichende ‚Fähigkeiten‘, um den Anforderungen zu entsprechen, die im Straßenverkehr an ihn gestellt werden.

usw.

(Teilfrage II) Was die **Einstellungen** angeht, die für das Fahrverhalten relevant sind, so bekundet Herr V. eine Denkweise, die frei ist von Risikotendenzen. Er bemüht sich um rationale Kontrolle seines Verhaltens im Straßenverkehr und will so den sozialen Aspekten gerecht werden.

usw.

(Alkoholproblematik) Schwieriger ist eine Urteilsbildung bei der **Alkoholproblematik**. Herr V. betrachtet die Verkehrsstraftat, die er unter Alkoholeinfluß beging, als eine Ausnahme. Für die Richtigkeit dieser Sichtweise spricht die lange unfallfreie Fahrgeschichte: Soweit erkennbar, hat er sich nach Alkoholkonsum von anderen Personen fahren lassen, statt sich selber ans Steuer zu setzen. Mitzubeachten ist auch die Tatsache, daß es ungewöhnliche Umstände waren, die zusammentrafen, als es zu der Fahrt nach Alkoholgenuß kam, bei der er von der Polizei gestellt wurde.

usw.

(„Empfehlung“ an die Behörde) Faßt man die einzelnen Aussagen zusammen, so ergibt sich: Aus einer Beurteilung der Leistungsfunktionen und einer Bewertung der Einstellungen lassen sich keine Bedenken ableiten gegen eine Wiedererteilung der Fahrerlaubnis.

Bedenken könnten bleiben angesichts der Alkoholproblematik. Die Diskussion der Einzelpunkte dürfte jedoch gezeigt haben, daß vieles gegen eine Deutung im Sinne einer Dauer-gefahr spricht.

Insgesamt betrachtet, erscheinen die Bedenken soweit entkräftet, daß Herrn V. die Fahrerlaubnis wiedererteilt werden kann.

Zweites Beispiel zur Stellungnahme: ***Gliederung nach drei Orientierungsfragen***

Auch das zweite Beispiel bleibt eine Skizze. Es geht um einen 39jährigen Probanden, der von seinem behandelnden Arzt zu einem Psychologen geschickt wurde. Eine Untersuchung sollte klären, warum der Klient unter ‚grundlosen‘ Ängsten leide, worauf seine psychosomatischen Beschwerden zurückgingen (Juckreize, Rückenschmerzen, Schlafstörungen) und wie es zu einer langanhaltenden Arbeitsunfähigkeit gekommen sei (Kasten 21-10).

Kasten 21-10: **Stellungnahme**

Fragestellung: Psychosomatische Beschwerden

Gliederung: Auch hier leitet der Gutachter die Stellungnahme mit einer Wiederholung der Ausgangsfrage ein.

Den weiteren Text gliedert er nach drei Fragen:

- Welche **Probleme** liegen vor? Er zählt die Probleme auf, die er als zentral einstuft. Als Konflikte macht er sie kenntlich, indem er ihre gegensätzlichen Pole nennt (z. B. einerseits Kontaktscheu, andererseits Kontaktwünsche).
- Auf welche **Bedingungen** gehen die Konflikte zurück? Die Probleme versucht der Gutachter zurückzuverfolgen bis zu ‚Erklärungen‘ und ‚Bedingungen‘ in der Biographie des Probanden.
- Lassen sich **Lösungsvorschläge** formulieren? Der Gutachter diskutiert Interventionsvorschläge, von denen er sich eine Besserung verspricht.

Hinweis: Wir fügen die drei Orientierungsfragen in Klammern zum Text hinzu. Im Original erscheinen sie nicht.

Text: Die psychologische Untersuchung sollte klären, warum Herr Walter P. unter grundlosen Ängsten und vielfaltigen körperlichen Beschwerden leidet und wie es zu der längeren Arbeitsunfähigkeit gekommen ist.

(Orientierungsfrage I: Welche Probleme liegen vor?)

Die Untersuchung hat unterschiedliche **Schwierigkeiten und Konflikte** sichtbar gemacht, die kurz charakterisiert seien:

- Herr P. leidet unter einer Unfähigkeit, Gefühle zu äußern, und wünscht zugleich lebhaft, seine Emotionen ausdrücken zu können,
- Er bekundet einen hohen Respekt vor seiner Frau (die er als ‚männlich‘ erlebt), bezeichnet sich jedoch gleichzeitig als unfähig, ihr seine Gefühle und Wünsche zu zeigen.
- In seinen Äußerungen gibt er lebhaft aggressive Impulse gegen seine Umwelt zu erkennen (vor allem gegen seine Geschwister, zwei Schwestern, und gegen seine Arbeitskollegen), zugleich aber beklagt er seine Unfähigkeit, seine Aggression auch offen zu zeigen.
- Ihn bestimmt eine starke Scheu vor Kontakten (bis hin zu gezielten Versuchen, sich selbst zu isolieren), aber ebenso ein starkes Streben nach enger Verbindung zu anderen Menschen, besonders zu nahestehenden Personen.

- Er erlebt sich als sehr selbstunsicher, er zeichnet von sich ein negativ getöntes Selbstbild, äußert aber gleichzeitig intensive Wünsche nach Anerkennung und Selbständigkeit.
- Er spricht von enger Verbundenheit zu seinen Eltern (von Gefühlen der Dankbarkeit und Schuldigkeit), daneben beschreibt er sein Empfinden, von den Eltern weder früher noch heute akzeptiert worden zu sein.
- Konflikte, so sagt er, kann er nicht durch eigene Aktivität bewältigen, sondern nur, indem er ausweicht und andere Personen um Hilfe bittet.

(Orientierungsfrage II: Worauf gehen die Probleme zurück?)

Auf dem Hintergrund solcher Konflikte werden die vielfältigen Ängste verständlich (mit jedem Konflikt ist Angst verbunden), ebenso die körperlichen Störungen als Ausdruck unverarbeiteter Probleme, schließlich auch die Arbeitsunfähigkeit als eine Art ‚Zusammenfassung‘ seiner Nöte und als Signal an die Umwelt.

Mit dieser Feststellung leiten wir zu dem Versuch über, die Probleme und Konflikte von Herrn W. in einen **Erklärungszusammenhang** zu setzen:

- Herr P. erlebt sich seit seiner Kindheit als nicht angenommen von den Eltern. (Seine beiden Schwestern seien immer bevorzugt worden.) Offensichtlich entwickelte er in dieser Umgebung ein zu geringes ‚Urvertrauen‘. Hier könnte seine Selbstunsicherheit ihre Wurzeln haben.
- In den Berichten des Probanden dominierten drei weibliche Autoritätspersonen (die Mutter und zwei ihrer Schwestern, die mit in der Familie lebten und sie ‚mitregierten‘), der Vater tritt kaum in Erscheinung. Wurde die männliche Geschlechtsrolle als zu wenig ‚reizvoll‘ erlebt? Jedenfalls gelang es Herrn W. nicht, sich mit der männlichen Rolle zu identifizieren, so daß er sich auch heute noch, wie er immer wieder betont, im Kontakt zu Frauen unsicher fühlt.
- Diese Konstellation, so scheint es, hat sich in der Ehe ‚wiederholt‘. Herr P. ‚suchte‘ sich eine ‚männliche‘ Ehefrau, ihr gegenüber will er demonstrativ Mann sein (so berichtet er selber), das aber mißlingt, und dies Ergebnis deutet er als ‚Versagen‘.
- Aus Selbstunsicherheit und mangelndem Selbstvertrauen neigt er dazu, vor Konflikten eher auszuweichen als sie aktiv zu bewältigen. Sicherheit sucht er in Abgelegenheit, sogar in Selbstisolation.
- Zwei Todesfälle im letzten Jahr (Tod des Vaters und des Schwiegervaters) und ein eigener Autounfall könnten Auslöser dafür gewesen sein, das eigene Unbehagen in körperlichen Beschwerden zu manifestieren - wobei angenommen wird, daß sich starke gefühlshafte Belastung in körperlichen ‚Zeichen‘ ausdrücken kann.

(Orientierungsfrage III: Was kann geschehen, um die Probleme zu lösen?)

Bei einem solchen Störungsbild ist eine psychotherapeutische **Intervention** einer Behandlung allein mit Medikamenten vorzuziehen.

- Anzueraten ist ein aufdeckendes Vorgehen, es sollte Herrn W. nicht ‚gestattet‘ werden, die ‚Ursachen‘ zu übersehen.
- Es empfiehlt sich eine individuelle Therapie, jedenfalls zu Beginn der Behandlung.
- Nächste Ziele sollten sein, die emotionale Blockierung zu lockern. Leisten könnten dies therapeutische Interventionen wie Gesprächs- oder Gestalttherapie.
- Weiteres Ziel könnte die Entwicklung eines ‚realitätsnäheren‘ Bezuges zur Mitwelt sein. Dazu wurden sich (später) eine Gruppentherapie eignen.

Insgesamt führen wir (somit) die Ängste und körperlichen Beschwerden von Herrn W. auf starke Konflikte zurück, die sich **in** und **aus** seiner Biographie entwickelt haben und die weitgehend seine gegenwärtige Situation bestimmen. Die Arbeitsunfähigkeit verstehen wir als Folge und Ausdruck dieser unbewältigten Probleme. Eine Hilfe halten wir für möglich, wenn sich Herr P. in eine längerfristige geplante psychotherapeutische Behandlung begibt.

21.5 Fehlertendenzen

Wir gehen auf zwei Arten von Fehlern ein, die eine Begutachtung verfälschen können.

Die **allgemeinen Fehler**, die bei der Verhaltensbeobachtung besprochen wurden (S. 199), können auch die Begutachtung beeinflussen, etwa die „zentrale Tendenz“, der „Hof-“ und „Positionseffekt“ - zusammenfaßbar unter dem Titel der „impliziten Persönlichkeitstheorie“. So kann ein Gutachten ganz unter dem „Ersten Eindruck“ formuliert und verfälscht werden. Auf diese Klasse von Fehlern sei nur verwiesen.

Für **spezielle Fehler**, die das Gutachten betreffen, nennt die Literatur vielfältige Beispiele (Fisseni, 1992, 120-127; Hartmann, 1973, 103-119; Schmidt, L.R., 1995, 476; Westhoff & Kluck, 1991, 139-149, 151-157; Zuschlag, 1992, 155-171).

Aufmerksam gemacht sei auf einige typische Fehler. In Anlehnung an Hartmann (1973, 103) seien zwei Klassen unterschieden:

1. sprachlich-stilistische Mängel,
2. sachlich-inhaltliche Fehler.

Zu 1.: Sprachlich-stilistische Mängel:

Hinweise auf stilistische Mängel dürften umstritten bleiben. Es gibt keine Regeln, die „vorschreiben“, welche sprachliche Gestalt ein Prosatext wie ein Gutachten annehmen „muß“.

- **Einhellig dürfte allerdings gelten:** Der größte Fehler besteht darin, ein Gutachten abzufassen ohne Blick auf den Empfänger. Es wäre ein Verstoß gegen ein zentrales Bestimmungsstück des Gutachtens - gegen das Anliegen einer „Kommunikation zwischen Experten und Laien“. Dieser Fehler liegt vor, wenn beispielsweise die Beschreibung der Verfahren, die Abfolge der Erkenntnissschritte, die Wortwahl und der Satzbau in einer „akademischen“ Fachsprache formuliert werden, die den Adressaten wie eine Fremdsprache anmuten muß.
- Den anderen Hinweisen zur Sprachgestalt sei eher der Rang „*bedenkenswerter Anregungen*“ zuerkannt:
 - ⇒ Fachtermini anzuwenden, sie aber dem Gutachten-Leser als Laien nicht zu erläutern stellt den Weg zum Verständnis. **Beispiel:** „*Im WIT, Gesamtergebnis, erreichte Frau Philippsen einen Prozentrang von 66.6.*“
 - ⇒ Ebenso schwer hat es der Empfänger, wenn ihm „schulgebundene“ Interpretationen angeboten, aber nicht „übersetzt“ werden. **Beispiel:** „*Die Probandin hat ihre Angst im Sinne eines Tribschicksals in ihr Gegenteil, in ein forsches Auftreten, verkehrt.*“

⇒ Schachtelsätze, zu lang gestreckt und unübersichtlich gebaut, sind dazu geeignet, im Leser Mißmut zu wecken und sein Unverständnis zu fördern - er „verirrt“ sich und verzichtet auf „Sinnentnahme“.

Hartmann zitiert ein köstliches *Beispiel* - an Skurrilität kaum zu überbieten (1973, 112): „*Die Intelligenz des Probanden, die, wenn man sie, was man bei der vorliegenden Fragestellung, die einen entsprechenden Vergleich fordert, zu tun nicht umhin kann, an der Norm, d. h. der Norm seiner in diesem Fall der studentischen, Bildungsgruppe bemißt, als, zumindest was bestimmte Einzelfähigkeiten, wie etwa den Umgang mit Zahlen, sei es beim praktisch-rechnerischen Denken, sei es bei Aufgaben, die abstrakte Kombination verlangen, angeht, eher unterdurchschnittlich bezeichnet werden muß, kann doch, besonders dann, wenn, wie es aufgrund der vielfältigen Interessen des Pb, die, wie bereits erwähnt, allerdings mehr emotionaler als sachlicher Natur sind, oft der Fall ist, sein Engagement angesprochen wird, besondere Leistungen bewirken.*“

⇒ Nicht viel besser wurde ein Gutachter verfahren, der nur kurze Stummelsätze aneinanderreihet. Bei Hartmann kann man für diesen „Athmastil“ Beispiele lesen (1973, 113).

⇒ Ein Fehler ist es auch, wenn Hauptgedanken in Nebensätze abgedrängt werden. Ein *Beispiel*, wieder von Hartmann (1973, 113): **Falsch:** „*Die Frage der Glaubwürdigkeit der Zeugin läßt sich dahingehend beantworten, daß sie wahrscheinlich einzuschränken ist.*“ **Richtig und einfacher:** „*Die Glaubwürdigkeit der Zeugin muß eingeschränkt werden.*“

Zu 2.: Sachlich-inhaltliche Fehler:

Eine Liste der Fehler, die sich auf die sachlichen und inhaltlichen Aspekte eines Gutachtens beziehen, muß ein offener Katalog bleiben, es lassen sich nur Beispiele anführen.

Zuerst seien Fehler erwähnt, die gegen den Gutachten-Aufbau verstoßen, wie er in diesem Kapitel vorgeschlagen wurde.

- Gegen das Konzept der *Übersicht* richten sich Verstöße wie die folgenden:
 1. Die Fragestellung wird zu breit geschildert.
 2. Informationsklassen werden nicht genannt, die später vorkommen.
 3. Verfahren werden aufgenommen, die mit der Fragestellung nichts zu tun haben.
- Gegen das Konzept der *Vorgeschichte* verstoßen bestimmte Darstellungsschritte:
 1. Die Informationsquellen werden nicht gekennzeichnet; damit wird eine Nachkontrolle erschwert.
 2. Die Wiedergabe der „Vorlagen“ wird zu breit ausgewalzt.
- Mit dem *Untersuchungsbericht*, wie er in unserem Schema vorgesehen ist, bleiben unvereinbar
 1. eine vorzeitige Festlegung ‚definitiver‘ Aussagen, ebenso
 2. eine ab- oder aufwertende Darstellung der Ergebnisse.

In den *Befund* gehören keine Aussagen, die Sachverhalte betreffen wie die folgenden:

1. Verhaltensweisen sind nur durch *ein* Verfahren belegt. (Es gibt Ausnahmen: Scores aus bewährten psychometrischen Verfahren oder spezielle „unerfindbare“ Details aus Gesprächen können auch als *Einzelbelege* auftreten!)
 2. Statt Verhaltensbeschreibungen tauchen diagnostische oder prognostische Erklärungen auf.
 3. Statt übergreifender (relativ konstanter) Verhaltensweisen werden Einzelbeobachtungen aus Vorgeschichte oder Untersuchungsbericht wiederholt.
 4. Das Gesamtverhalten des Probanden wird von einigen zentralen „Dominanten“ her „strukturiert“ - die eher einen „Idealtypus“ entwerfen als die Persönlichkeit des Probanden auch mit ihren unaufgehellten Widersprüchen zu schildern.
- Gegen das Konzept der *Stellungnahme* verstoßen Vorgehensweisen wie die folgenden:
1. Diagnose, Prognose oder Entscheidungsvorschlag überschreiten den Umfang der Informationen, die in Vorgeschichte und Untersuchungsbericht vorgegeben und im Befund „gebündelt“ werden.
 2. Die „Antwort“ auf die Fragestellung wird parteiisch formuliert.
 3. Der urteilende Psychologe versucht andere als psychologische Sachverhalte (etwa juristische Fragen) zu klären.

Nun seien einige Fehler angeführt, die eher übergreifender Natur sind, sich jedenfalls nicht auf einen bestimmten Gutachten-Abschnitt beziehen:

- Der Aufbau eines Gutachtens ist so unklar, daß er für den Empfänger bei der Lektüre nicht transparent wird.
- Verfehlt sind Aussagen, die geeignet sind, im Probanden ablehnende Einstellungen zu verstärken, beispielsweise Anflüge von Hilfs- und Hoffnungslosigkeit. **Beispiel:** *Der fünfzehnjährige Dietmar hat einen Ladendiebstahl begangen. Das Gutachten zieht die „Folgerung“: „Es ist damit zu rechnen, daß Dietmar seinen Hang zur Kleptomanie nicht meistern lernt“ (Schmidt, L.R., 1995, 476).*
- Ebenso falsch sind Aussagen, die negative Befunde „entschuldigen“, „schön reden“, „banalisieren“. **Beispiel:** *Herr Schmitz ist angeklagt seine Vermieterin, Frau Werfel, tätlich angegriffen und gewürgt zu haben. Im Gutachten heißt es dazu: „Zwar ist das Verhalten des Angeklagten zu mißbilligen, aber seine Aggressionen werden durchaus verständlich, wenn man die schäbige Art berücksichtigt, mit der Frau Werfel Herrn Schmitz seit über zwei Jahren schikaniert hat.“*
- Zur Klärung der Fragestellung tragen Aussagen nicht bei, die so vage gefaßt sind, daß sie fast auf Jedermann zutreffen. **Beispiel:** *„Eine gewisse Bindungsfähigkeit ist vorhanden, jedoch sind auch Wünsche nach Unabhängigkeit unverkennbar“ (Schmidt, L. R., 1995, 476).*

- Wenn in Vorgeschichte oder Untersuchungsbericht Verhaltensweisen registriert werden, die (nur) in einem bestimmten situativen Kontext auftreten, ist es ein grober Verstoß, diese zu generalisieren. **Beispiele:** *Herr X erzählt von seiner Hundephobie. Falsch wäre dann der Satz „Herr X ist leicht erregbar und überaus ängstlich.“ - Der Schüler Holger Y erzählt, daß er sich bei der letzten Deutscharbeit nicht konzentrieren konnte. Falsch wäre die Generalisierung: „In der Schule kann sich Holger Y nicht konzentrieren“ (Schmidt, 1995, 476).*

Unsere Darstellung berücksichtigt Fehler, die der Gutachter selber begehen (oder meiden) kann. Außer Betracht bleiben die Irrtümer, die der Empfänger begeht, indem er das Gutachten falsch interpretiert - falsch aus der Sicht des Gutachters (Graumann, 1967; Pelzmann, 1972; Haas, 1975). Diese Fehlerquelle kann der Gutachter nur kontrollieren, wenn er sein Gutachten mündlich vorträgt und an der Interaktion mit dem Adressaten merkt, ob seine Aussagen in dem beabsichtigten Sinne aufgenommen werden.

Konsequenz: Ein Gutachten, das völlig frei von Fehlern wäre, gibt es wohl nur in Lehrbüchern - vermutlich aber auch in Lehrbüchern nicht, sondern nur in den Köpfen von Lehrbuch-Autoren. Hinweise auf Fehler zielen darum darauf ab, den Gutachter zu steter Selbstkritik zu ermuntern, und dazu, sich unablässig einer kollegialen Kontrolle zu unterziehen.

21.6 Zusammenfassung zu Kapitel 21

Eine wichtige Aufgabe psychologischer Tätigkeit, die Begutachtung, wurde als Beispiel integrativer multimodaler Diagnostik vorgestellt. Da psychologische Begutachtung auf höchst unterschiedlichen Tätigkeitsfeldern zu leisten ist, kann niemand den Gutachtern eine einheitliche Vorgehensweise vorschreiben.

Wenn in diesem Kapitel dennoch ein bestimmtes Gutachtenschema vorgeschlagen wurde, dann aus der Absicht, dem Anfänger eine Möglichkeit der Übung und des Trainings zu geben. Die Leitidee, an der sich dieses Grundschema orientiert, ist die des Colloquiums zwischen dem Gutachter als Experten und dem Fragesteller als Laien.

An diesem Ziel orientieren sich sowohl die Gliederung als auch die Sprache des Gutachtens, wie sie hier vorgeschlagen werden. Die Gliederung sieht fünf Abschnitte vor: Übersicht, Vorgeschichte, Untersuchungsbericht, Befund, Stellungnahme. Die Sprache vermeidet Parteinahme, sie liefert Erklärungen statt Wertungen, sie ist abgestimmt auf das Sprachverständnis des Gutachten-Empfängers.

Der Gutachter sollte die Fehler kennen, die seine Urteile verzerren könnten.

21.7 Kontrollfragen zu Kapitel 21

- Umschreibung des gutachterlichen Prozesses.
- Sozial-ethischer Kontext des gutachterlichen Prozesses.
- Möglicher Aufbau eines Gutachtens.
- Rolle der ‚Übersicht‘.
- Rolle der ‚Vorgeschichte‘.
- Rolle des ‚Untersuchungsberichtes‘.
- Rolle des ‚Befundes‘.
- Rolle der ‚Stellungnahme‘.

22. Kapitel

Beispiel III Integrativer Diagnostik: *Beurteilung von Stellenbewerbern*

Als drittes Beispiel integrativer Diagnostik skizzieren wir Beurteilung und Auswahl eines Stellen-Bewerbers - ein Vorgehen, das sich der diagnostischen Aufgabe einer *Selektion* zuordnen läßt.

Die Darstellung gliedert sich in sieben Abschnitte:

- Vorausgesetzte Situation und Aufgabenstellung (22.1),
- Erfassung der Stellenanforderungen (22.2),
- Vorauswahl der Bewerber (22.3),
- Beurteilung der ausgewählten Bewerber (22.4),
- Unterrichtung abgelehnter Kandidaten (22.5),
- Evaluation, Erfolgskontrolle (22.6),
- Ethische Implikationen einer Bewerberbeurteilung (22.7).

Es folgen eine Zusammenfassung (22.8) und eine Reihe von Kontrollfragen (22.9).

22.1 Vorausgesetzte Situation und Aufgabenstellung

Vorausgesetzt sei folgende **Situation**: Ein Unternehmer sucht für eine Stelle einen ‚angemessenen‘ Kandidaten, etwa für die Position eines Leiters der Abteilung ‚Kundenbeziehungen‘. Er wendet sich an ein Personalberatungsinstitut und bittet um Suche und Beurteilung von Bewerbern.

Der Urteilsgang wird sich um zwei Probleme zentrieren:

- Ein Kandidat muß gesucht und *beurteilt* werden. Einen Vorrang erhält dabei die Aufgabe, zu klären, wie die Tätigkeit aussieht, für die ein Kandidat gesucht wird: *Soll-Profil*, und wie die Person ausgestattet ist, die sich um die Stelle bewirbt: *Ist-Profil*.
- Suche und Beurteilung müssen *dem Auftraggeber mitgeteilt* werden. Diese *Rückmeldung* muß sicherstellen, *erstens* daß die Konstrukte der Kandidatenbeurteilung das Stellen- und das Bewerberprofil angemessen ‚abbilden‘, und *zweitens* daß Untersucher und Auftraggeber die Konstrukte in dem gleichen Sinne verstehen (Kleinevoss, 1978, 13).

Somit erfordern Kandidatenbeurteilung und Rückmeldung eine Erkenntnisbemühung, die Aspekte der Arbeit mit Aspekten der Persönlichkeit verbindet (Hoyos, 1986, 60; Funke & Schuler, 1986, 36; Trebeck, 1961).

Im einfachsten Falle können Bewerber-Beurteilung und Rückmeldung an den Auftraggeber einander als zwei zeitlich getrennte Prozesse folgen (zuerst Beurteilung, dann Rückmeldung). In komplexeren Fällen ist eine Sequenz vorzuziehen, in der Urteil und Rückmeldung einander mehrfach ablösen. Die folgende Darstellung orientiert sich an diesem zweiten Modell.

22.2 Erfassung der Stellenanforderungen

Die Anforderungen stammen zwar vom Auftraggeber, aber er gibt sie nicht ausformuliert vor (aufgelistet wie in einem Katalog), sondern eingeschlossen in das Stellenangebot, das er macht. Erste Aufgabe des Gutachters ist es, die Position genauer zu charakterisieren und Anforderungsfacetten zu artikulieren - ein Urteilsprozeß, der sich in einzelne Abschnitte aufgliedert (Bach, 1986, 109; Hoyos, 1986, 61; Jochmann, 1984, 119-120; Maukisch, 1978, 124).

Es liegt nahe, zunächst das *Umfeld der Stelle* ‚aufzunehmen‘ (Frieling & Sonntag, 1987, 136-152). Festzuhalten sind etwa: Produktpalette des Unternehmens / Vorteils- und Nachteilspositionen in der Branche / Branchenkonventionen / Managementphilosophie / Führungsstruktur, Führungsstil / Betriebsklima / usw.

Anschließend sollte sich eine genaue *Stellenbeschreibung* (Sauermann, 1979, 29-34) mit Angaben wie: Stellenbezeichnung / Zuordnung hinsichtlich Funktion (Finanzen, Verkauf, Produktion usw.) und Hierarchie (Sachbearbeiter, Gruppenleiter, Abteilungsleiter usw.) / Aufgabenbereiche im Rahmen der Unternehmensziele: so speziell wie möglich / Besondere Befugnisse / Stellvertretung / usw.

Eine Beschreibung des Umfeldes und der Stelle erlauben es, die *Qualifikationsmerkmale* zu formulieren, die von einem Bewerber erwartet werden, etwa: Art des Ausbildungsganges / Erwünschter Abschluß (z. B. Akademischer Grad) / Vertrautheit mit speziellen Aufgaben (Führungserfahrung, Prokura usw.) / Besondere Kenntnisse, die erwünscht sind (z. B. Beherrschung von Fremdsprachen) / Verhaltensmerkmale, die erwünscht oder unerwünscht sind / Alter (Obergrenze, Untergrenze) / usw.

Aus Umfeld- und Stellenbeschreibung sollte sich auch das *Anforderungsprofil* ‚ableiten‘ lassen: Als Soll-Angabe benennt es ‚Persönlichkeitsmerkmale‘, die bei dem Stellenbewerber zu fordern sind (Frieling & Sonntag, 1987, 153; Kompa, 1984, 59). Gewünschte Ausprägungsgrade kann der Beurteiler auch quantitativ ‚abbilden‘ (Jochmann, 1987, 2). Enthalten könnte ein solches Profil

Dimensionen wie: Intellektuelles und kognitives Potential / Verbale Gewandtheit / Entscheidungsfähigkeit / Durchsetzungsvermögen / Führungskompetenz / Leistungsmotivation / Leistungsverhalten / Soziale Kompetenz / Frustrationstoleranz / usw.

In diesem Kontext stellen sich dem Psychologen zwei Aufgaben:

- Auswahl von Methoden zur Erfassung des Stellenprofils,
- Rückmeldung des Stellenprofils an den Auftraggeber.

Auswahl von Methoden zur Erfassung des Stellenprofils

Zur Erfassung des Anforderungsprofils (des Soll-Profiles) stehen unterschiedliche Methoden und Techniken zur Verfügung, zum Teil begründet auf theoretischen Ansätzen, in der Mehrzahl jedoch pragmatischer Natur (Frei, 1981, 23; Frieling, 1975, 37; Frieling, 1977, 33-57).

Es seien fünf Beispiele angeführt:

- **Dokumenten- und Inhaltsanalyse:** Mit Hilfe einer Inhaltsanalyse werden vorhandene ‚Dokumente‘ im Blick auf die Stelle, die zu besetzen ist, aufgeschlüsselt. In Frage kommt jeder Text, der Aussagen zu der Stelle macht: Leitsätze zur Personal- und Unternehmensführung ebenso wie Dienstvorschriften oder Unternehmenspläne, Arbeitsplatz- und Tätigkeitsbeschreibungen ebenso wie Ausbildungsberichte, Mitarbeiterbeurteilungen oder Publikationen wie Werbeschriften oder Statistiken usw.
- **Interview:** Ein Interview kann viele Informationen liefern, dazu verschiedenartige Bereiche abdecken. Bei Verständnisschwierigkeiten sind Erklärungen möglich. Verbunden mit solchen Gesprächen sind allerdings jene Vor- und Nachteile, welche die unterschiedlichen Interviewformen kennzeichnen: Standardisierte Interviews erschließen nur vorgegebene Informationsklassen, unstandardisierte Interviews erschweren den Vergleich der erhobenen Informationen (Kap. 8, S. 215).
- **Fragebogen:** Gemeint ist ein Fragebogen in Analogie zu einem Persönlichkeitstest, der es ermöglicht, vielerlei Informationen in standardisierter Form zu sammeln (z.B. der ‚Fragebogen zur Arbeitsanalyse (FAA)‘ von Frieling & Hoyos, 1978).
- **Checklisten:** Eine Checkliste besteht aus einem Katalog von Feststellungen, die Stellenanforderungen umschreiben. ‚Anzukreuzen‘ ist, wie weit eine Anforderung (ein ‚Item‘) jeweils auf eine bestimmte einzelne Stelle zutrifft. Ankreuzen kann jeder, der mit der Stelle vertraut ist: Vorgesetzter, Mitarbeiter oder Personalfachmann (Bach, 1986, 109).
- **Methode der kritischen Ereignisse:** Im Rahmen der ‚Arbeitsanalyse‘ bezeichnen ‚kritische Ereignisse‘ (critical incidents) solche Verhaltensmuster, die erforderlich sind, damit eine Arbeit effektiv, ökonomisch und sinnvoll ausfällt. Drei Schritte sind vorgesehen:

1. Die kritischen Ereignisse werden gesammelt. Befragt werden Stelleninhaber, Experten, Vorgesetzte. Angestrebt wird eine Kooperation aller Beteiligten (vgl. Bach, 1986, 109).
2. Die Ereignisse werden skaliert nach dem Grad, in dem sie eine Tätigkeit erleichtern oder vervollkommen. Beteiligt werden Experten.
3. Eine Liste wird angelegt, die jene Verhaltensweisen enthält, welche einen guten Stelleninhaber charakterisieren.

Resümee: Jede Methode bietet Chancen, setzt aber auch Grenzen. Sie im Individualfall gegeneinander abzuwägen gehört zur Kunstfertigkeit des urteilenden Psychologen.

Dabei ist festzuhalten (Weinert, 1981, 187): Je höher eine Leitungsfunktion, deren Anforderungen formuliert werden sollen, desto weniger eignen sich dazu Methoden, die auf Generalisierung und Quantifizierung abheben.

Rückmeldung des Stellenprofils an den Auftraggeber

Die Angaben zu den Anforderungen sammelt der Beurteiler in Kooperation mit dem Auftraggeber und bündelt sie in einem Stellenprofil. Über das ausformulierte Profil sollte er dem Auftraggeber eine Rückmeldung geben, zumindest mündlich, am besten jedoch, zur Vermeidung von Mißverständnissen, auch schriftlich; denn in dieser Phase lassen sich Mißverständnisse noch ausräumen. Wechselseitig können Beurteiler und Auftraggeber noch einmal festlegen, was die einzelnen Facetten des Profils konkret besagen. *-Die Auftragsbestätigung sollte das (korrigierte) Stellenprofil einschließen.*

22.3 Vorauswahl der Bewerber

Dem Anforderungsprofil ist das Bewerberprofil gegenüberzustellen. Es stellen sich drei Aufgaben:

- Rekrutierung einer Zielgruppe von Bewerbern,
- Erstellung einer Rangreihe der Bewerber,
- Rückmeldung der Rangreihe an den Auftraggeber.

Rekrutierung einer Zielgruppe von Bewerbern

Diese Zielgruppe kann prinzipiell auf zwei Wegen rekrutiert werden:

- aus **innerbetrieblichem** Potential (z. B. durch Qualifizierung, Umschulung, Versetzung von Mitarbeitern);
- aus **außerbetrieblichem** Potential (z. B. durch Auswertung von Stellengesuchen, durch vielfältige Formen von Werbung).

Die Argumente für beide Wege sind im konkreten Falle gegeneinander abzuwägen:

- Für interne Rekrutierung spricht die Möglichkeit der **Personalentwicklung**: Wenn der Arbeitsmarkt keine Bewerber ‚freigibt‘, können durch Umschulung aus betriebseigenem Personal Stellenanwärter ‚kreiert‘ werden.
- Für externe Personalbeschaffung spricht vor allem ein Argument: Externe regen eher zu **Innovationen** an.

Für Spitzenpositionen ist ein bedeutsamer Weg der Werbung die **Anzeige** in einer Zeitung (Mason & Belt, 1986). Neben der Auswahl des ‚richtigen‘ Blattes sind dabei insbesondere zu bedenken: die Art der Blätter, in denen eine Anzeige erscheinen soll (Niveau, Leserkreis, Leserprofil) / Größe der Anzeige / Umfang der Angaben aus dem vorliegenden Stellen- und aus dem erwünschten Bewerberprofil / usw.

Zu formulieren ist sodann der Katalog von **Angaben**, die **bei einer Bewerbung** erwartet werden, beispielsweise: Persönliche Daten / Bildungsweg (Schul-, Studienzeugnisse, Diplome) / Berufliche Ausbildung (Ausbildungs-, Arbeitszeugnisse) / Berufliche Erfahrungen / Beruflicher Werdegang (Firma, Branche) / Besondere Kenntnisse / Besondere Interessen / Referenzen / Berufliche Erwartungen / Gehalts- oder Vergütungsvorstellungen / usw. (Frieling & Sonntag, 1987, 75).

Erstellung einer Rangreihe der Bewerber

Die Akten und Dokumente, welche die Bewerber einsenden, sind nach verschiedenen Kriterien zu sortieren. Zuerst werden die Schreiben mit einem Sperrvermerk ausgesondert, dann die anderen gesichtet. Dabei orientiert sich der Sortiervorgang letztlich an der Korrespondenz von Anforderungs- und Bewerberprofil, soweit sich diese Korrespondenz in den Unterlagen abbildet.

Die Bewerber werden in eine Rangreihe gebracht, beispielsweise mit den Klassen:

- ‚Ideale‘ Bewerber („Sofort einladen!“);
- Bewerber, die für die Stelle in Betracht kommen, aber (leichte) Schwächen aufweisen;
- Bewerber mit auffälligeren Schwächen;
- Bewerber mit erheblichen Schwächen („Kommen nur im Ausnahmefall für eine Bewerbung in Betracht“);
- Bewerber, die für eine weitere Stufe der Auswahl nicht in Betracht kommen.

¹ *Sperrvermerk* bezieht sich auf folgenden Sachverhalt: Der Betrieb, der eine Stelle ausschreibt, kann anonym bleiben; in der Ausschreibung erscheint dann nur eine Chiffre. Sollte ein Bewerber **dem** Betrieb angehören, der unter der Chiffre erscheint, kann er verlangen, daß sein Schreiben nicht geöffnet wird - er notiert dann auf seinem Schreiben einen „Sperrvermerk“.

Der Beurteiler kann auch versuchen, die Kandidaten unter verschiedenen Gesichtspunkten zu klassifizieren, also unterschiedliche Rangreihen zu bilden.

Rückmeldung der Rangreihe an den Auftraggeber

über die ersten beiden Gruppen (Idealbewerber, Bewerber mit leichten Schwächen) sollte eine Rückmeldung an den Auftraggeber gehen - gegebenenfalls nur in mündlichem Vortrag: Dies wurde die Vertraulichkeit der Bewerber-Informationen länger wahren als eine schriftliche Mitteilung unter Namensnennung. Mit dem Auftraggeber zusammen wird entschieden, welche Bewerber zu einer eingehenden Beurteilung eingeladen werden.

22.4 Beurteilung der ausgewählten Bewerber

Die Dimensionen für die Kandidatenbeurteilung liefert das Anforderungsprofil. Ob die erwünschten Merkmale bei einem Bewerber gegeben sind, soll der weitere Urteilsprozeß ermitteln.

Die Darstellung des Urteilsprozesses sei gegliedert wie folgt:

- Einige Voraussetzungen der Bewerberbeurteilung (22.4.1),
- Einzelschritte der Bewerberbeurteilung (22.4.2),
- Rückmeldung der Bewerberbeurteilung (22.4.3),
- Selbstpräsentation des Bewerbers vor dem Auftraggeber (22.4.4),
- Auswahlentscheidung (22.4.5),
- Nachbearbeitung (22.4.6).

22.4.1 Einige Voraussetzungen der Bewerberbeurteilung

Bevor die Hauptschritte einer Bewerberbeurteilung geschildert werden, seien in einem Überblick einige Voraussetzungen genannt:

- **Ziele der Kandidatenbeurteilung:** Dem Auftraggeber kann daran liegen, von dem Bewerber nur eine *Fähigkeitsbeschreibung* (ein individuelles Leistungsprofil) oder aber ein *Persönlichkeitsbild* zu erhalten, das als einen wesentlichen Teil die Fähigkeitsbeschreibung enthält.
- **Kriterienfrage:** Zu vergleichen sind Anforderungs- und Kandidatenprofil, sie dienen als Kriterien - eine einfache Prozedur, wenn beide Profile exakt formuliert sind) erst recht, wenn sie darüber hinaus objektiv quantifiziert und hinreichend reliabel und valide sind (vgl. Jochmann, 1984, 119-120, 126-129; Kompa, 1984, 48-95). Genau an diesem Punkte können erhebliche Schwierigkeiten auftreten - die sich zentrieren um die Frage, wie im Einzelfall die Meßqualität der Kriterien zu bewerten ist.

Fehlerquellen beim Beurteiler: Der Diagnostiker sollte damit rechnen, daß seine ‚Voreingenommenheit‘ und seine ‚Erwartungen‘ die Suche und die Verarbeitung von Informationen **auch** verzerren können.

Fehlerquellen beim Beurteilten: Wie der Diagnostiker, so kann auch der Bewerber Informationen verfälschen. So kann der Proband Sachverhalte verschweigen oder verschleiern. - Wie soll sich der Psychologe verhalten? Er kann

- ⇒ nach *Lücken im Lebenslauf* forschen,
- ⇒ nach *Widersprüchen* in den Aussagen suchen,
- ⇒ den Bewerber immer wieder ‚nötigen‘, *konkretes Verhalten* zu schildern (das sich meist nicht so geschickt verfälschen läßt).

- **Marketing-Situation als Fehlerquelle:** Der Diagnostiker ist auch Kaufmann, der seine Dienste ‚verkaufen‘ will. Um den Auftraggeber nicht als Kunden zu verlieren, könnte er sich dazu verführen lassen, einen ‚weniger geeigneten‘ Kandidaten zu empfehlen - wenn unter den Bewerbern kein ‚geeigneter‘ Kandidat auszumachen ist.

22.4.2 Einzelschritte der Bewerberbeurteilung

Wenn der Untersucher sich über Ziele und Kriterien klargeworden ist, die Situation und die Art der Urteilsbildung festgelegt hat, dann kann er die geeigneten Methoden auswählen. Er verknüpft Tätigkeits- und Personenmerkmale, indem er die Erfassung bestimmter Merkmale bestimmten Verfahren zuordnet.

Nun sind Personenmerkmale nicht in Relation zu bestimmten Managementfunktionen definiert worden: Weder Intelligenz noch Konzentration, weder Extra- / Introversion noch sensurnotorische Reaktion wurden beispielsweise im Hinblick auf Führungspositionen konzipiert. Demnach gilt: „Für die Entscheidung, mit welchen Verfahren welche Merkmale zu messen sind, existieren keine eindeutigen Regeln“ (Kompa 184, 114; vgl. Durchholz, 1981, 276).

Darum muß der Untersucher im Blick auf die Fragestellung klären, ob ein bestimmtes Instrument im Einzelfall das gesuchte Merkmal erfaßt.

Heranziehen kann er dafür alle Instrumente, welche die Psychodiagnostik anbietet. Genannt seien hier einige Verfahrensklassen (Hornthal, 1985 a, b; Kompa, 1984, 114, 159; Lehrenkrauss, 1986; Lippert & Zeidler, 1986; Loretto, 1986):

- *Interview(techniken),*
- *standardisierte Verfahren: (Intelligenz- und andere Leistungstests /übliche Persönlichkeitstests, speziell biographische Fragebögen usw.),*
- *unstandardisierte Verfahren: (Fallanalysen, Gruppendiskussionen, Rollenspiele usw.),*
- *Verhaltensbeobachtung*
- *u s w .*

Der Gutachter sollte *multimodal* vorgehen, also unterschiedliche Instrumente einsetzen, die sich wechselseitig stützen. - *Projektive Verfahren dürften dem Anliegen nicht affin sein, sich darum verbieten.*

Die *Untersuchungssequenz* sollte so angelegt sein, daß sie den Kandidaten für die Mitarbeit gewinnt. Darum empfiehlt sich eine Reihenfolge der Verfahren, die einen plausiblen Zusammenhang zur ‚Eignung‘ herstellt.

Letztlich kommt es darauf an, daß der *Gutachter* sich selbst zu einem *sensiblen subjektiven ‚Meßinstrument‘* ausbildet. Die objektiven Instrumente darf er nicht abwerten. Doch steht und fällt die Qualität der Kandidatenbeurteilung zuerst mit der Erfahrung und Kunstfertigkeit des urteilenden Psychologen.

22.4.3 Rückmeldung der Bewerberbeurteilung

Das Ergebnis wird in einem Bericht an den Auftraggeber zusammengefaßt.

Vier Leitideen

Die Rückmeldung sollte vier formalen Leitideen entsprechen (Dirks, 1961, 615); die Mitteilung sollte

- **vollständig** sein (alle Konstrukte enthalten, welche für die ausgeschriebene Funktion oder Position relevant sind);
- **prägnant** sein (die Konstrukte eindeutig beschreiben, Überschneidungen, also Redundanz, vermeiden);
- **strukturiert** sein (den Zusammenhang sichtbar machen, in dem die Konstrukte zueinander stehen);
- **kommunikationsfreundlich** (dem Auftraggeber ohne umständliche Erklärungen verstehbar).

Grad der Standardisierung

Die Rückmeldung kann unterschiedliche Grade von Standardisierung annehmen:

- **Standardisiert** ist eine Rückmeldung, wenn die Urteilsdimensionen und die Art ihrer Wiedergabe festgelegt sind.
- **Halbstandardisiert** ist eine Rückmeldung, wenn zwar bestimmte Vorgaben gemacht werden, ihre Ausgestaltung jedoch offen bleibt.
- **Unstandardisiert** ist eine Rückmeldung, wenn dem Untersucher völlig freie Hand bleibt, sowohl bei der Auswahl der Urteilsdimensionen als auch bei der Art ihrer Beschreibung. Maßstab der Darstellung ist allein die Regel, die Kandidatenbeurteilung so zu gestalten, daß die Antwort vollständig und verständlich ausfällt.

Insgesamt dürfte gelten: Je idiosynkratischer eine Position ist, um die sich ein Kandidat bewirbt, um so weiter wird sich die Beurteilung von einer Standardfassung entfernen.

Schriftliche oder mündliche Mitteilung

Der Gutachter kann sein Urteil schriftlich mitteilen oder mündlich vortragen.

Eine **schriftliche Mitteilung** hat mehrere Vorteile. Der Auftraggeber kann die Urteilsschritte im Zusammenhang betrachten, er kann immer wieder die Schlüssigkeit der Argumente überprüfen.

Nachteil: *Ein Gutachten, das nur schriftlich angeboten wird, kann nicht sicherstellen, daß ein „diagnostisches Urteil in dem gemeinten Sinne“ aufgenommen wird (Graumann, 1967, 124).*

Solche Klarheit und Eindeutigkeit zu ermöglichen ist größter Vorzug einer **mündlichen Vermittlung**. In Wechselrede zwischen Psychologen und Auftraggeber können Urteilskategorien erläutert, Bewertungen kommentiert, Mißverständnisse aufgeklärt werden.

Nachteil: *Aussagen, die nur mündlich vorgetragen werden, haben den Charakter des ‚Flüchtigen‘, sie können nicht beliebig reproduziert und erneut kontrolliert werden.*

Darum spricht vieles für eine **Kombination von schriftlicher Fassung und mündlichem Kommentar**, wo immer sie möglich ist. Auf diese Weise lassen sich die Sorgfalt schriftlicher Fixierung und die Flexibilität mündlicher Darstellung verbinden.

Umfang der Darstellung

Um ihrem Ziel gerechtzuwerden, sollte die Rückmeldung Angaben wie die folgenden enthalten:

- eine präzise Formulierung der Fragestellung,
- den Namen des Auftraggebers,
- den Namen des Untersuchers,
- den Namen des Bewerbers,
- Damm, Dauer und Ort der Untersuchung,
- wichtige Informationen aus der ‚Vorgeschichte‘ des Kandidaten,
- die Liste der Auswahlkriterien,
- die Liste der Informationsquellen und der Untersuchungsverfahren,
- eine Darstellung der relevanten Ergebnisse,
- eine Beschreibung der Stärken und Schwächen des Kandidaten,
- Gehalts- / Vergütungsvorstellungen des Kandidaten,
- Laufbahnvorstellungen des Kandidaten,
- Name und Unterschrift des verantwortlich zeichnenden Untersuchers.

Empfehlung an den Auftraggeber?

Die gesamte Rückmeldung müßte zulaufen auf eine Empfehlung - *wenn* denn der Gutachter eine Empfehlung geben will.

Wenn der Auftraggeber nicht ausdrücklich eine Empfehlung fordert, ist dem Diagnostiker zu raten, bei einer *deskriptiven Diagnostik* zu bleiben (Stoll, 1977, 207). Dies bedeutet, daß er möglichst verhaltensnah die Stärken und Schwächen des Kandidaten beschreibt und sie gegeneinander abwägt.

Übrigens drückt eine deskriptive Darstellung prägnant die Rolle des urteilenden Psychologen aus: Der Gutachter hat die Rolle eines *Entscheidungs-Helfers*, nicht die eines *Entscheidungs-Trägers*.

22.4.4 Selbstpräsentation des Bewerbers vor dem Auftraggeber

Aus der schriftlichen Begutachtung erfährt der Auftraggeber Grundlegendes über die Person des Kandidaten. Je bedeutender aber die ausgeschriebene Position ist, desto größeren Wert wird er darauf legen, den Bewerber auch persönlich zu beurteilen, er wird wünschen, daß sich der Kandidat ihm selber vorstellt. In diesem Falle übernimmt der Psychologe die Rolle eines Moderators.

Diese Präsentation kann neue Gesichtspunkte der Bewertung erbringen, beim Kandidaten ebenso wie beim Auftraggeber. Denn wieder laufen zwei Urteilstgänge ab: Der Auftraggeber beurteilt den Kandidaten, der Kandidat den Auftraggeber.

22.4.5 Auswahlentscheidung

Die Aufgabe, die jetzt ansteht, kann nicht der Psychologe, sie kann nur der Auftraggeber lösen: Der Auftraggeber entscheidet, welcher Kandidat *ausgewählt* wird. Der Gutachter hat Gesichtspunkte beigetragen, die dem Auftraggeber diese Entscheidung ermöglichen sollen.

22.4.6 Nachbearbeitung

Den Psychologen können weitere Aufgaben erwarten, etwa, im Sinne einer ‚Nachbereitung‘, die Bitten oder Aufträge wie die folgenden einschließen kann:

- ergänzende Referenzen einzuholen und zu bewerten,
- ergänzende Gespräche mit dem Kandidaten zu führen,
- den Arbeitsvertrag, wenn erwünscht, mitzubespochen.

Gerade diese Phase dürfte belegen, daß die Auswahl von Bewerbern auch *Kompromißbereitschaft* erfordert.

22.5 Unterrichtung abgelehnter Kandidaten

Zum Urteilsprozeß gehört auch die Entscheidung, auf welche Weise *abgelehnte* Kandidaten unterrichtet werden (Fisseni, 1995; Maier, 1980, 49-57; Stoll, 1977, 209-213). Dazu zwei Überlegungen:

1. Es trägt zur kritischen Selbsteinschätzung des **Kandidaten** bei, in diesem Sinne auch zu seinem Wohlbefinden, wenn er die Gründe erfährt, die zu seiner Ablehnung geführt haben (und somit nicht auf Vermutungen oder Spekulationen angewiesen ist).
2. Es kann das Image der **Psychologen** und der Personalberatung, der er angehört, verbessern, wenn der abgelehnte Kandidat in einer Weise unterrichtet wird, die es ihm ermöglicht, seinen Beurteiler oder einen anderen Psychologen als Personalberater erneut zu akzeptieren, wenn es um eine neue Bewerbung geht.

22.6 Evaluation, Erfolgskontrolle

Letztes Kriterium für die Validität der Beurteilung ist der Erfolg, der dem ausgewählten Kandidat auf der ausgeschriebenen Position zuerkannt wird (Funke, U., 1986, 92; Heneman, 1986, 815).

Darum gehört es zur Aufgabe des Diagnostikers, nach einem angemessenen Zeitraum festzustellen, ob die getroffene Wahl ‚richtig‘ war - festzustellen, ob der eingestellte Kandidat seine Aufgaben erfolgreich bewältigt.

*Bei dieser Evaluation ist mit dem **Fehler** zu rechnen, daß der ‚Prophet sich selbst bestätigt‘ (self-fulfilling prophecy). Sowohl der Auftraggeber der einen Bewerber eingestellt hat, wie auch der Psychologe, der mitgewirkt hat, sind daran interessiert, zu ‚beweisen‘, daß ihre Entscheidung richtig war:*

Sind jedoch beide Seiten hinreichend selbstkritisch - daran interessiert, Schwachstellen ihrer Entscheidungskette zu entdecken, dann werden sie sich auch problematische oder falsche Entscheidungen eingestehen.

Einen ‚Erfolg‘ zu bewerten, erfordert Kriterien. Kriterien für ‚Erfolg auf einer Position‘ sind nicht ein für alle Male definiert. Als Beispiele seien genannt: Beförderung, Gehaltszuwachs, Leistungsbeurteilung durch Vorgesetzte usw. Jedes dieser Kriterien schließt eigene Probleme ein (Kompas, 1984, 50; Lippert & Zeidler, 1986; Maukisch, 1986, 86-87).

22.7 Ethische Implikationen einer Bewerberbeurteilung

Abgeschlossen sei dieser Beitrag mit einem Hinweis auf ethische Implikationen. Dazu nur Stichworte:

- Respekt vor der Personenwürde des Probanden (keine herabsetzende Formulierung),
- Informierung des Probanden über Weitergabe von Daten,
- Schutz der Privat- und Intimsphäre des Probanden bei der Weitergabe von Informationen,
- Schweigepflicht gegenüber (unbefugten) Dritten,
- Datenschutz (sichere Aufbewahrung relevanter, Vernichtung überholter Daten).

Die Beachtung ethischer Implikationen dürfte bei Kandidaten und Auftraggebern die Akzeptanz von Eignungsuntersuchungen erhöhen.

22.8 Zusammenfassung zu Kapitel 22

Als Beispiel von Selektion wurde ein spezieller Fall geschildert: die Beurteilung von Kandidaten, die sich um eine Führungsposition bewerben.

Der Urteilsgang zentrierte sich um zwei Aufgaben:

1. Ein Kandidat muß gesucht und beurteilt werden.
2. Das Urteil muß dem Auftraggeber vermittelt werden.

Die Stellenanforderungen müssen erfaßt und in einem Stellenprofil formuliert werden. Eine Zielgruppe von Bewerbern muß rekrutiert, aus ihr müssen Kandidaten ausgewählt und beurteilt, ihre Qualifikation in einem Bewerberprofil zusammengefaßt, schließlich Stellen- und Bewerberprofil verglichen werden.

über Teilergebnisse und Gesamtergebnis ergehen Rückmeldungen an den Auftraggeber. Die Auswahlentscheidung liegt bei dem Auftraggeber.

Die abgelehnten Bewerber sollten angemessen unterrichtet, die ‚Richtigkeit‘ der Auswahl überprüft, die ethischen Implikationen beachtet werden.

22.9 Kontrollfragen zu Kapitel 22

- Grundzüge der Aufgabe.
- Instrumente zur Erfassung der Stellenanforderungen.
- Instrumente zur Erfassung der Bewerbereignung.
- Probleme beider Instrumentenklassen.
- Fehlerquellen der Bewerberbeurteilung.

-
- Rückmeldungen an den Auftraggeber.
 - Umfang der Rückmeldung.
 - Ethische Implikationen der Bewerberbeurteilung.

23. Kapitel

Beispiel IV Integrativer Diagnostik: *Assessment-Center (AC)*

Als viertes Beispiel integrativer Diagnostik sei das Assessment-Center (AC) skizziert. Dieses Verfahren lässt sich interpretieren als eine Aufgabe, die sowohl Selektion wie auch Klassifikationen einschließt. - Das Assessment-Center dient beispielsweise einer Bewerberselektion, die sich unter anderem an ‚gutachterlichen‘ Zuordnungen, also an Klassifikationen, orientiert. - Ein Assessment-Center kann auch anderen Zielen dienen, beispielsweise der Personalentwicklung oder dem Training von Mitarbeitern (Jeserich, 1990, 36).

Kapitel 23 behandelt folgende Themen:

- Abgrenzung (Definition) (23.1),
- Zeitliche Konzeption (23.2),
- Urteilsdimensionen (23.3),
- Übungen (23.4),
- Vorrang der Verhaltensbeobachtung (23.5),
- Ablaufbeispiel (23.6),
- Auswertung (23.7),
- Validität (23.8).

Es folgen eine Zusammenfassung (23.9) und eine Reihe von Kontrollfragen (23.10).

23.1 Abgrenzung (Definition)

An einem Assessment-Center sind in der Regel **drei Personengruppen** beteiligt (Fisseni & Fennekels, 1995, 23):

- Die Personen, die beurteilt werden sollen, heißen *Teilnehmer*.
- Die Personen, welche die Teilnehmer beurteilen, heißen *Beobachter*.
- Beide Gruppen, Beobachter wie Teilnehmer, stehen unter der Leitung eines sogenannten *Moderators*.

Das Anliegen besteht darin, eine Anforderungssituation, (etwa die Aufgaben einer Führungsposition) in einer ‚Testsituation‘ möglichst realistisch abzubilden.

den. *Das Testverhalten soll demnach eine realistische Stichprobe des Anforderungsverhaltens repräsentieren.*

Daraus leiten sich **vier Konstruktionsprinzipien** ab:

Anforderungsnähe: Die Aufgaben, welche die Teilnehmer zu lösen haben, sollen den Anforderungen der Zielposition ähnlich sein.

Verhaltensnähe: Die Leistungen, die das Assessment-Center den Teilnehmern abfordert, sollen verhaltensnah definiert sein, damit sie eindeutig zu beobachten sind. - Darum kann man auch von Beobachtungsnähe sprechen.

Verfahrensvielfalt: Die Methoden, mit denen die Leistungen der Teilnehmer erfaßt werden, sollen vielfältig (multimodal) sein, damit Artefakte vermieden werden.

Beobachtervielfalt: An der Beurteilung der Teilnehmer sollen sich mehrere Beobachter beteiligen, damit Fehlertendenzen sich ausgleichen.

Die Aufgaben, welche die Teilnehmer lösen sollen, werden zusammengestellt in sogenannten **Übungen**; sie tragen unterschiedliche Namen, beispielsweise ‚Gruppendiskussion‘ oder ‚Planspiel‘ oder ‚Präsentation‘. In den Übungen soll sich die „Anforderungsnähe“ des Assessment-Centers darstellen, weil sie jenes Verhalten evozieren sollen, das in der Zielposition erwartet wird (Fisseni & Fennekels, 1995, 23).

Eine Umschreibung (Definition), die auf viele Assessment-Centers zutreffen dürfte, gibt Kasten 23-1.

Kasten 23-1:
Definition des Assessment-Center
Quelle: Fennekels (1987, 10)

„Das Assessment-Center Verfahren ist
ein systematisches und flexibles Verfahren
zur
kontrollierten und qualifizierten Feststellung
von Verhaltensleistungen und -defiziten,
das von mehreren Beobachtern
gleichzeitig
für mehrere Teilnehmer
in bezug auf vorher festgelegte Übungen
und
bestimmte Anforderungen
vornehmlich zur
Personalauswahl und -Weiterentwicklung
von vielen Personalentwicklungsabteilungen
in Großunternehmen
mit Erfolg und steigender Tendenz eingesetzt wird.“

Einige charakteristische Stichworte seien kommentiert:

- Das Assessment-Center ist **ein systematisches und flexibles Verfahren:**

- ⇒ *Systematisch*: Die Verhaltensleistungen und Verhaltensdefizite der Teilnehmer sollen unter strenger Kontrolle erfaßt werden.
- ⇒ *Flexibel*: Das Vorgehen soll auf die ‚individuelle Situation‘ einer Organisation, etwa eines Unternehmens, zugeschnitten werden und in diesem Sinne flexibel sein. (Die Flexibilität reicht so weit, daß sich die Methode sogar auf einen einzelnen Bewerber abstimmen läßt: Einzel-Assessment, Jochmann, 1987; Schmid, 1995).
- Das Assessment-Center sieht **mehrere Beobachter und mehrere Teilnehmer** vor:
 - ⇒ *Mehrere Beobachter*: Die Verhaltensfassung beruht nicht auf der diagnostischen Urteilskraft eines einzelnen, sondern auf einem Verbund kognitiver Kapazität - in der Erwartung, daß die Stärken des einen Beobachters die Schwächen des anderen ausgleichen. Die ‚vereinigte Kapazität‘ kann allerdings nur dann diagnostisch wirksam werden, wenn die Beobachter angemessen ‚trainiert‘ worden sind.
Die Beobachter sind in der Mehrzahl keine Fachpsychologen, sondern gehören „dem oberen Managementkreis an, in der Regel sind sie zwei Hierarchiestufen höher angesiedelt als die Bewerber“ (Frieling & Sonntag, 1987, 78).
 - ⇒ *Mehrere Teilnehmer*: Auf diese Weise ist es möglich, die Qualitäten der Teilnehmer voneinander abzuheben durch wechselseitigen Vergleich (Figur-Grund-Bezug).
Als Faustregel gilt, daß etwa zwölf Teilnehmer von sechs Beobachtern beurteilt werden.
- Das Assessment-Center beruht auf **Übungen, die vorher festgelegt** werden: Sie sollten „maßgeschneidert“ sein für die Organisation, in deren Dienst ein Assessment-Center geplant wird.
AC-Anforderung & AC-Übungen, die für einen bestimmten Betrieb konzipiert wurden, sind im Idealfall idiosynkratisch, sie lassen sich nicht unverändert übertragen auf einen anderen Betrieb.
- Ziel** eines Assessment-Centers ist **Personalauswahl und Personalentwicklung**. In diesem Sinne handelt es sich um eine Selektionsaufgabe, die klassifikatorische Mittel mit-einsetzt.
 - ⇒ *Personalauswahl* bezeichnet Maßnahmen, die dazu bestimmt sind, die Eignung von Personen für Berufe, Stellen oder Tätigkeiten zu erfassen (Dorsch, 1994, 556).
 - ⇒ *Personalentwicklung* bezeichnet Maßnahmen zur Analyse, Planung, Förderung und Evaluation des personellen Potentials einer Organisation; Ziel ist es, die Effizienz zu verbessern und so die personellen Ressourcen eines Unternehmens zu erkennen und zu fördern (Dorsch, 1994, 558).
- Eingesetzt wird das Assessment-Center vor allem **in Großunternehmen**; denn, wie die Beschreibung erkennen läßt, „erfordert der Einsatz solch eines Instrumentes erhebliche organisatorische, administrative und methodische Vorleistungen sowohl auf seiten eines Assessment-Center-Konstruk-

teurs als auch für den Anwender. Das macht das Verfahren zeit- und kostenintensiv und bis heute primär für große bis mittelgroße Organisationen interessant“ (Fennekels, 1987, 11).

23.2 Zeitliche Konzeption

Ein schematischer Überblick soll die zeitliche Konzeption eines Assessment-Centers veranschaulichen (Jeserich, 1990, 35). Drei Phasen sind vorgesehen: Vorbereitung, Durchführung und Abschluß.

1. *Vorbereitung:*

- Festlegen der Ziele und der Zielgruppe,
- Auswahl der Beobachter,
- Definition des Anforderungsprofils (ggf. mit den Beobachtern),
- Zusammenstellen der Übungen im Blick auf die Anforderungen,
- Information der Teilnehmer.

2. *Durchführung:*

- Training der Beobachter,
- Empfang der Teilnehmer: Erläuterung von Ziel und Ablauf des Programms,
- Bearbeiten der Übungen und Unterlagen durch die Teilnehmer,
- Beobachten der Leistungen durch die Beobachter,
- Auswertung der Beobachtungen.

3. *Abschluß und Rückmeldung:*

- Abstimmen der Auswertungen,
- Anfertigen der Gutachten/ Empfehlung von Maßnahmen,
- Endabstimmung/Endauswahl,
- Informierung der Teilnehmer über Ergebnisse,
- Vereinbarung von Förder- oder Entwicklungsmaßnahmen.

23.3 Urteilsdimensionen

Das Assessment-Center soll herauszufinden helfen, wie weit die Teilnehmer in der Lage sind, Aufgaben zu bewältigen, die an eine vorgesehene Zielposition gebunden sind.

Der Wunsch eines Unternehmens besteht darin, „Personen mit dem entsprechenden Potential herauszufinden. Dieses Potential zu definieren erweist sich als ziemlich schwierig, da Managementpositionen selbst innerhalb eines Unternehmens sehr unterschiedlich sein können“ (Frieling & Sonntag, 1987, 77).

Zwar sollen Aufgaben und Anforderungen individuell auf einzelne Unternehmen abgestimmt werden. Aber es gibt Merkmale, die in den Anforderungsprofilen vieler Assessment-Centers auftauchen (Fermekels, 1987, 36; Frieling & Sonntag, 1987, 79; Hess & Schmitt-Planert, 1985, 189; Jochmann, 1987; Schuler & Stehle, 1987).

Exemplarisch seien einige Anforderungsdimensionen genannt:

- *Administrative Fähigkeiten:* Organisations- und Planungsfähigkeit, Entscheidungskraft und Verantwortungsbewußtsein;
- *Soziale Kompetenz:* Empathie, Sachlichkeit, Durchsetzungs-, Überzeugungskraft, Flexibilität, Loyalität;
- *kognitive Kompetenz:* Intelligenzhöhe, analytisches Denken, systematisches Denken, Sprachbeherrschung;
- *Leistungsverhalten:* Konzentration, Ausdauer, Streßresistenz, Frustrationstoleranz;
- *Selbstbild:* Selbstbewußtsein, emotionale Selbstkontrolle, Wertung der eigenen beruflichen Vergangenheit, Erwartung für die berufliche Zukunft, Aufstiegsorientierung
- u s w .

23.4 Übungen im Assessment-Center

Alle psychologischen Methoden und Aufgaben, welche inhaltlich die erwünschten Anforderungen abbilden, eignen sich dazu, die gesuchten Dimensionen zu erfassen. Zusammengestellt werden diese Aufgaben und Methoden in den Übungen.

Es seien einige Beispiele angeführt (Fennekels, 1987; Fisseni & Fermekels, 1995, 53-73; Jeserich, 1990; Jochmann, 1987; Lehrenkrauss, 1986):

Gruppendiskussionen

Ein Anlaß wird vorgegeben, über den die Teilnehmer (führerlos oder unter Leitung) diskutieren müssen.

Beispiele:

1. *Jeder Teilnehmer soll begründen, warum der Dienstwagen, der neu angeschafft worden ist, vor allem ihm zur Verfügung stehen sollte.*
2. *Jeder Teilnehmer soll erklären, welche Probleme entstehen, wenn ein Mitarbeiter zum Vorgesetzten seiner ehemaligen Kollegen ernannt wird.*

Präsentation

Jeder Teilnehmer muß vor dem Kreis der Beobachter ein Thema vorstellen. Eine gewisse Vorbereitungszeit wird eingeräumt.

Fallbeispiel

Vorgegeben werden Situationen im Rahmen konkreter Organisationsstrukturen. Ein Problem, etwa ein Konflikt zwischen Betriebsmitgliedern, wird geschildert. In Einzelarbeit soll der Teilnehmer eine Lösung vorschlagen und rechtfertigen.

Beispiel (Jeserich, 1990, 157): „Als der Personalleiter durchs Werk geht und zwischen die Stanzerei und Fräseerei kommt, fliegt ihm auf einmal ein Knäuel Putzwolle in den Nacken. Er dreht sich sofort um, kann aber nicht mehr feststellen, aus welcher Richtung die Putzwolle geworfen worden ist. Beide Meisterabteilungen bieten das Bild normal arbeitender Arbeitsgruppen. Keiner der Arbeitenden scheint vom Personalleiter Notiz zu nehmen. - Wie sollte sich der Personalleiter verhalten? Bitte begründen Sie Ihre Vorschläge.“

Was interessiert, ist nicht nur der Inhalt der Replik, sondern vor allem auch die Art der Begründung.

Postkorb

Dem Teilnehmer werden Posteingänge, Notizen, Anfragen usw. vorgelegt, die sich in Dringlichkeit, Komplexität und Bedeutsamkeit für die Firma erheblich unterscheiden. „Die Aufgabe besteht darin, alle Dokumente zu lesen und dann Entscheidungen zu treffen“ (Fisseni & Fennekels, 1995, 64).

Den Postkorb bearbeitet jeder Teilnehmer in Einzelarbeit innerhalb einer vorgegebenen Zeitspanne.

Rollenspiel

Den Teilnehmern werden bestimmte Rollen zugewiesen, etwa die von Leitern verschiedener Abteilungen. Die Übung ermöglicht viele Varianten.

Beispiele:

1. Konflikte zwischen Vorgesetzten und Untergebenen sollen ausgetragen und gelöst werden. Von je zwei Teilnehmern soll einmal der eine, ein andermal der andere die Rolle des Vorgesetzten und des Untergebenen übernehmen.
2. Ein Beobachter spielt den Gruppenleiter ‚Fracht‘ der Lufthansa. Von ihm soll der Teilnehmer „als Vertreter einer renomierten Wirtschaftszeitung mit entsprechenden Rahmeninformationen einen Anzeigenauftrag akquirieren“ (Jochmann, 1987, 2).

Explorationen

Jeder Teilnehmer erhält Gelegenheit, seine privaten und beruflichen Erfahrungen zu schildern und seine Karrierevorstellungen zu entwickeln. „Er kann weitere Ergänzungen und Erklärungen zu seinem bisherigen Verhalten im Assessment-Center geben und seine gefühlsmäßige Situation darin beschreiben“ (Hess & Schmitt-Planert, 1986, 190).

Leistungs- und Persönlichkeitstests

In manchen Assessment-Centers werden übliche Leistungs- und Persönlichkeitstests mitverwandt. Ein solches Vorgehen wirft jedoch spezielle Probleme auf. Der ‚Arbeitskreis AC‘ zählt den Schritt sogar zu den ‚Verstößen‘ gegen die Standards (1992, 2).

Warum? Das Assessment-Center beruht nach seiner Konzeption auf der Analyse konkreter Verhaltensprozesse in einer Organisation.

„Diesem Ansatz widersprechen Leistungs- und Persönlichkeitstests nach ihrer Konzeption; denn sie beziehen sich auf generelle und generalisierbare Verhaltensweisen... Diese Generalität von Leistungs- und Persönlichkeitstests widerspricht der Spezifität des Assessment-Centers.

Mit anderen Worten: Zwischen den generellen Merkmalen, die ein Leistungs- oder Persönlichkeitstest erfaßt, und den spezifischen Anforderungen, auf die ein Assessment-Center abhebt, besteht eine große Distanz“ (Fisseni & Fennekels, 1995, 70).

23.5 Vorrang der Verhaltensbeobachtung

Während der Übungen sammeln die Beobachter ihre Informationen über die Teilnehmer. Vorrang erhält dabei die Verhaltensbeobachtung; denn Verhaltensbeobachtung ist, wie der ‚Arbeitskreis Assessment-Center‘ mit Nachdruck betont, das zentrale Instrument der Datenerhebung (1992, 2).

Dieser Vorrang folgt aus der Tatsache, daß im Assessment-Center Verhaltensweisen realisiert werden und Verhaltensbeobachtung sich definiert durch die Funktion, in gezielten Wahrnehmungen Verhalten in seiner Performanz zu erfassen (Fisseni & Fennekels, 1995, 19).

Verhaltensbeobachtung bezeichnet vor allem Fremdbeobachtung. Dabei soll die Fremdbeobachtung *eines* Beobachters gestützt werden durch die Fremdbeobachtung *anderer* Beobachter. Darum sieht das Assessment-Center den Einsatz *mehrerer* Beobachter vor.

„Um die Eigenart der Verhaltensbeobachtung zu wahren, gilt für das Assessment-Center der Rat, Beobachtungen von Urteilen zu trennen. Beobachtungen sind unmittelbare Ergebnisse der Wahrnehmung. Urteile dagegen sind Schlußfolgerungen aus solchen unmittelbaren Wahrnehmungen“ (Fisseni & Fennekels, 1995, 19) .

23.6 Ablaufbeispiel

Wie Konstruktion und Auswahl der Übungen ist auch der Ablauf flexibel gestaltbar. Darum können nur Exempel angeführt werden. Hier ein Beispiel nach Frieling und Sonntag (1987, 78-79).

Montag

Vormittag:

- *Einführung*: Erläuterung der Aufgaben, Vorstellung der Beobachter, Information über Ziele;
- *Erste Gruppendiskussion* der Teilnehmer über ein frei wählbares Thema;
- *Fallbearbeitung*: Stellungnahme zu einem Problem, das schriftlich vorgegeben wird und schriftlich zu bearbeiten ist.

Nachmittag:

- *Managementaufgabe*: Im Rahmen von Haushaltsverhandlungen soll jeder Teilnehmer für seinen Bereich ein möglichst hohes Budget ‚herausholen‘.

Abend:

- *Verkaufsaufgabe*: Für den Verkauf eines Produktes soll jeder Teilnehmer einen Verkaufsplan ausarbeiten, abgestimmt auf das jeweilige Unternehmen.

Dienstag

Vormittag:

- *Präsentationsübung*: Jeder Teilnehmer stellt einem ‚Vorgesetzten‘ seinen Verkaufsplan vor (den er am Montagabend ausgearbeitet hat).
- *Rollenspiel*: Ein Teilnehmer führt, unter der Kontrolle eines Beobachters, mit einem anderen Teilnehmer ein Entlassungsgespräch.
- *Postkorb*: Dem Teilnehmer wird eine Reihe von Briefen und oder Mitteilungen vorgelegt, „schriftliche Vorgänge, die für die entsprechende Zielposition möglichst typische Problemstellungen enthalten“ (Hess & Schmitt-Planer?, 1985, 189). Die Aufgaben, die sie stellen, müssen in eine Lösungsreihe gebracht, Wichtiges vor weniger Wichtigem plazierte werden: Termine, Absprachen mit anderen Firmen, interne Besprechungen, Absprachen mit städtischen Behörden, Delegation von Aufgaben usw.

Nachmittag:

- *Präsentation*: In einem Dialog mit einem ‚Vorgesetzten‘ (Beobachter) muß jeder Teilnehmer seinen Plan zur Erledigung der Postkorb-Aufgaben erläutern und vertreten.
- *Zweite Gruppendiskussion*: Die Teilnehmer sollen als Stadtverordnete ein Budget von acht Millionen aufteilen; jeder Teilnehmer vertritt dabei ein bestimmtes Ressort.

Mittwoch, Donnerstag

Unter Leitung des Moderators besprechen die Beobachter die Daten und Informationen, die sie bei den Übungen gesammelt haben, und verarbeiten sie zu Gutachten. Diese Gutachten bespricht und diskutiert je ein Beobachter mit je einem Teilnehmer.

23.7 Auswertung

Über die Teilnehmer liegen am Ende Informationen aus drei Quellen vor:

- Beobachter urteilen über die Teilnehmer (Beobachterurteile).
- Jeder Teilnehmer urteilt über jeden anderen Teilnehmer (Fremdurteile).
- Jeder Teilnehmer beurteilt auch sich selber (Selbstbeurteilungen).

Die Informationsquellen liefern Auskünfte

- aus verschiedenen Übungen (z. B. Gruppendiskussion, Postkorb),
- über verschiedene Anforderungsdimensionen (z. B. administrative, soziale, kognitive Kompetenz),
- von verschiedenen Beurteilern.

Die Informationen werden sowohl numerisch abgebildet (in Ratings) wie auch verbal gefaßt (in kurzen verhaltensnahen Notizen). *Den verbalen Urteilen kommt im Assessment-Center jedoch ein Vorrang zu*, weil sie differenzierter gefaßt werden können und in diesem Sinne aussagekräftiger sind.

Über jeden Teilnehmer wird am Ende ein ‚Gutachten‘ erstellt. Dieses Gutachten beruht auf den Informationen aller Beobachter, es „orientiert sich an der Leistung im Assessment-Center, und nur an ihr - nicht auch an Informationen, die einem Beobachter aus anderen Quellen zufließen, etwa aus Zeugnissen oder Referenzen. Schildern soll das End-Gutachten das Potential eines Teilnehmers: die Stärken, die er gezeigt hat, und die Schwächen, die zum Vorschein kamen“ (Fisseni & Fennekels, 1995, 139).

Rückmeldung: „Der Moderator sorgt dafür, daß die Teilnehmer spätestens innerhalb einer Woche nach Abschluß des Assessment-Centers eine Rückmeldung erhalten: Ein Beobachter spricht mit ihnen über die Stärken und Schwächen, die sie im Assessment-Center gezeigt haben, und gibt Empfehlungen, die besagen, wie sie ihre Stärken noch verbessern und ihre Schwächen korrigieren können“ (Fisseni & Fennekels, 1995, 31).

23.8 Validität

Das Assessment-Center soll eine Zielposition möglichst realistisch und möglichst individuell abbilden. Darin liegt seine Stärke, aber offensichtlich auch seine Schwäche: Je ‚individueller‘ die Abbildung, desto schwieriger eine Validierung.

Darüber, was das Assessment-Center leistet, sind sich die Autoren nicht einig.

Die Problematik ergibt sich bei der Entscheidung, welche Erfolgskriterien anerkannt werden (Maukisch, 1986, 86): Beförderung, Gehaltszuwachs, Leistungsbeurteilung (durch Vorgesetzte) usw.? Könnte sich aber dabei nicht auch erweisen, daß „das Assessment-Center nur das gängige Führungsrollenstereotyp in den Organisationen zementiert, weil seine Validität sich mehr auf korrelierte Kognitionen der Beteiligten hinsichtlich guten Führungsverhaltens, also auf ein gemeinsames Stereotyp bezieht, als auf einen Zusammenhang zwischen Verhaltensweisen im Assessment-Center und objektiv produktiven Ergebnissen der Tätigkeit im Unternehmen“ (Maukisch, 1986, 87)?

Sehr scharf, wohl zu hart, urteilen Frieling und Sonntag (1987, 79): „Die Perfektion psychologischer Eignungsdiagnostik garantiert zwar eine weitgehende Gleichbehandlung und Objektivität der Auswahl, sie degradiert den Bewerber jedoch zu einem Objekt, dem man geschickt seinen Widerstand bei der ‚Vermessung‘ abringt. Und bei all dem bleibt auch heute noch unklar, ob die Meßkriterien tatsächlich so valide sind, wie man das unterstellt.“

Andere urteilen zuversichtlicher, beispielsweise Funke, U. und Schuler (1986, 37): „Mehr-Rater-Methoden erweisen sich allgemein dem Einzelurteil als überlegen.“ Ähnlich sagt Hinrichs (1978, 597), daß sich das Assessment-Center (vielen Einwänden zum Trotz) als ein vergleichsweise verlässlicher Prädiktor erwiesen habe.

Vorsichtige Zustimmung äußert Maukisch (1986, 89): „Zusammenfassend und pauschal betrachtet läßt sich sagen, daß die ACS (Assessment-Center-Systeme) in ihrer Vorhersagekraft für Beförderungskriterien und subjektive Leistungskriterien alternativen Prädiktoren eher überlegen, zumindest aber gleichwertig erscheinen.“

Allerdings hält er es für wichtig, die Assessment-Centers weiterzuentwickeln (1986, 91). Die Untersucher sollten

- die Redundanz von Verfahren minimieren („Abspeckung der Mammutprogramme“),
- die Prüfaufgaben im Sinne einer Prozeßdiagnostik modifizieren (sie ‚dynamisieren‘),
- die Teilnehmer nicht mehr miteinander vergleichen, vielmehr ihre Leistung an der Aufgabenstellung messen.

23.9 Zusammenfassung zu Kapitel 23

Das Assessment-Center wurde beschrieben als ein verhaltens- und anforderungsnahes Verfahren im Dienste von Personalauswahl und Personalentwicklung. Die Anforderungen einer Zielposition sollen möglichst realistisch in den AC-Aufgaben abgebildet werden.

Die sogenannten Übungen sollen die Anforderungen so verhaltensnah abbilden, so daß alle Informationen einer Beobachtung zugänglich sind: Bei der Datenerhebung hat die Verhaltensbeobachtung einen Vorrang.

Beteiligt sind drei Personengruppen:

1. Beurteilt werden die Teilnehmer.
2. Es urteilen die Beobachter.
3. Beide Gruppen stehen unter der Leitung eines Moderators.

Meist sind mehrere Tage für die ‚Beobachtung‘ der Teilnehmer vorgesehen. Über jeden Teilnehmer wird am Ende ein Gutachten erstellt, das nur auf AC-Informationen beruht. Mit jedem Teilnehmer soll dieses Gutachten individuell besprochen werden.

Ein zentrales Problem erwächst aus der Aufgabe, die Urteile, die dem Assessment-Center entstammen, an Erfolgskriterien zu validieren.

23.10 Kontrollfragen zu Kapitel 23

Mehrfache Ziele.

Vier Charakteristika.

Rolle der Verhaltensbeobachtung.

Beteiligte Personengruppen.

Arten und Beispiele von Übungen.

Urteilsdimensionen.

Ablaufphasen.

Drei Datenquellen.

Probleme der Validierung.

Nachwort¹

Ratschlag an den Leser

Dieses Lehrbuch wurde für Studierende geschrieben, konzipiert von der ‚diagnostischen Situation‘ her. Immer wieder wurde gefragt: Welche Kenntnisse und Fertigkeiten sollte sich ein Psychologe angeeignet haben, wenn er sich auf eine diagnostische Situation einläßt? Das Buch sollte ihm dafür eine Grundlage vermitteln.

Diese Grundlage sollte der Studierende aber erweitern. In welche Richtung kann er dabei gehen?

Das Lehrbuch bietet ihm eine eher ‚formale Diagnostik‘. Ergänzen könnte sie eine stärker inhaltlich orientierte Diagnostik. Was ist damit gemeint?

Daß dieses Buch eine eher ‚formale Diagnostik‘ entwickelt, soll besagen: Es beschreibt, auf welche Weise der Diagnostiker eine Fragestellung aufschlüsseln und wie er eine Antwort suchen kann. Es gibt aber nicht an, wie er bestimmte Einzelmerkmale umschreiben, mit welchen Verfahren er sie erfassen kann. Beispielsweise leitet es ihn zwar an, wie er Fragebogen oder Explorationen handhaben solle, etwa zur Erfassung von Motivation; es sagt ihm jedoch nicht, was er unter Einzelmerkmalen wie ‚Angst‘ oder ‚Aggression‘ oder ‚Zurechnungsfähigkeit‘ zu verstehen habe und mit welchen Methoden er sie erfassen könne.

Diagnostisch relevante Merkmale im einzelnen zu beschreiben und einzelne Verfahren zu ihrer Erfassung zu benennen wäre Gegenstand einer eigenen Variante von Diagnostik. Eine solche ‚inhaltlich orientierte Diagnostik‘ liegt jedoch nicht ausformuliert vor.

Der Versuch, eine solche Variante zu entwerfen, wäre auch mit vielen Risiken behaftet; denn sowohl die Umschreibung wie auch die Erfassung von Einzelmerkmalen sind in hohem Maße klient- und situationszentriert. Der Einzelfall

¹ Diese konkreten Hinweise stammen von Herrn Diplompsychologen Dr. Dieter Vennen, einem langjährigen Mitarbeiter in unserer Abteilung „Diagnostik und Persönlichkeitspsychologie“.

bestimmt in hohem Maße mit, welche ‚Inhalte‘ in einem Merkmal wie ‚Angst‘ oder ‚Aggression‘ oder ‚Zurechnungsfähigkeit‘ enthalten sind.

Es gibt aber Bereiche, in denen sich der Studierende einer ‚inhaltlichen Diagnostik‘ nähern kann.

Diagnostik wird auf unterschiedlichen **Aufgabenfeldern** betrieben, etwa der Forensischen, der Klinischen, der Pädagogischen Psychologie, der Verkehrs-, Arbeits- und Organisationspsychologie. Diagnostische Aufgaben werden hier auch inhaltlich umrissen.

Der Studierende sollte versuchen, solche unterschiedlichen Fragestellungen und das zugehörige Instrumentarium kennenzulernen. Dabei wird er eine Auswahl treffen müssen, zum einen nach dem Angebot seines Studienortes, zum anderen nach dem Arbeitsfeld, das er später anstrebt.

Diagnostik wird von unterschiedlichen psychologischen **‚Schulen‘** her mitgeprägt. Unter verschiedenen Perspektiven werden gleiche diagnostische Aufgaben inhaltlich unterschiedlich formuliert.

Der Studierende sollte versuchen zu verstehen, wie Fragestellungen und Lösungswege aussehen, wenn er unterschiedliche Perspektiven übernimmt, beispielsweise so gegensätzliche Sichtweisen wie die einer ‚analytischen‘ oder einer ‚verhaltenstheoretischen‘ Schule. In dem unterschiedlichen Kontext nehmen sich Merkmale wie ‚Leistungsmotivation‘ oder ‚elterlicher Erziehungsstil‘ höchst unterschiedlich aus. Höchst unterschiedlich fällt darum auch die Wahl der Instrumente aus, die zu ihrer Erfassung benannt werden.

Einen weiteren Weg, ‚diagnostische Kenntnisse‘ zu verbreitem, eröffnet die **Lektüre**, beispielsweise von Literatur zu Gebieten wie den folgenden:

- Querverbindungen der Diagnostik zu anderen psychologischen und nicht-psychologischen Disziplinen,
- Anwendungsfelder psychologischer Diagnostik,
- Relevanz konkreter ethischer Postulate in der Psychodiagnostik,
- juristische Vorschriften, die für Diagnostiker wichtig sind,
- Vertiefung methodischer Grundlagen,
- Einzelfall-, Gruppen-, Familiendiagnostik
- usw.

Gezielt kann der Studierende auch **anwendungsorientierte Übungen und Seminare** besuchen mit dem Ziel, ‚zu lernen durch Tun‘. Beispiele:

- Einübung explorativer Techniken,
- Beachtung unterschiedlicher Formen nonverbaler Kommunikation,
- Aneignung und Anwendung eines Kanons von Leistungs- und Persönlichkeitstests im einzelnen,
- Aneignung und Einübung projektiver Verfahren im einzelnen,
- Training integrativer Techniken, beispielsweise der psychologischen Begutachtung,

- Reanalysen abgeschlossener Einzelfälle,
- Entwurf von Untersuchungsplänen für ‚neue‘ Einzelfälle
- usw.

Was in Seminaren und Übungen vorgestellt wird, läßt sich ergänzen durch individuelle Übungen ‚zuhause‘. Im **Selbstversuch** kann der Studierende beispielsweise Tests ausprobieren oder mit Bekannten Einführungsgespräche trainieren. Solche individuellen Übungen kann er ‚prüfen‘ lassen durch Kommilitonen, die Psychologen sind wie er, gelegentlich wohl auch durch einen Dozenten.

Darüber hinaus sollte er sich bemühen, solche **Praktika** zu finden, in denen erfahrene Psychologen ihn einführen ‚in die Praxis‘, somit seine Arbeit einer ‚Supervision‘ unterstellen und ihn an diagnostischen Aufgaben beteiligen (Tests und Fragebögen vorlegen, projektive Verfahren anwenden, ihre Auswertung mitbesprechen, Daten für Gutachten miterheben, Klienten bei der Rehabilitation mitbetreuen usw.). Findet er solche ‚sinnvollen‘ Praktika, könnte er sie, sofern möglich, freiwillig weiterführen.

Bei all diesen Bemühungen - der Wahl von Praktika ebenso wie bei Besuch von Seminaren oder bei Selbstversuchen - sollte er **Schwerpunkte** setzen, wiederum je nach dem Angebot seines Studienortes und nach dem später erwünschten Arbeitsfeld.

Gleichgültig jedoch, welche Aufgabe er später anzielt: In drei Bereichen sollte er seine Ausbildung vertiefen:

- Vertraut machen sollte er sich, durch Übung und Training, mit *explorativen Techniken* und ihrer theoretischen Einbettung.
- Beherrschen sollte er, dank Training und Anwendung, einen *Kanon von Leistungs- und Persönlichkeitstests*, er sollte ihren theoretischen Hintergrund kennen, zum Beispiel ihre Fundierung in Testtheorie und Persönlichkeitspsychologie.
- Auskennen sollte er sich mit den unterschiedlichen *Schritten einer integrativen Diagnostik*, beispielsweise mit Formen psychologischer Beratung, mit der Assessment-Center-Methode oder der psychologischer Begutachtung.

Spezielle Schwerpunkte kann er setzen je nach dem Arbeitsfeld, dem er sich zuwenden will. Dafür nur Fingerzeige:

- Wenn er in den Bereich der *Schulpsychologie* gehen will, sollte er das große Feld der Entwicklungs- und Schultests überschauen, er sollte sich einen umfassenden Überblick verschaffen über die Entwicklungspsychologie als ‚Psychologie der gesamten Lebensspanne‘, er müßte sich auch spezielle Rechtskenntnisse erwerben.
- Wenn er in das Arbeitsfeld der *Industriellen Psychologie* strebt, müßte er sich mit einem breiten Umfeld vertraut machen, beispielsweise mit Verfahren zur Erfassung von Stellenanforderungen, mit Methoden der Beur-

teilung, der Auslese und der Entwicklung von Personal. Auskennen müßte er sich auch in Ausschnitten der Nachbardisziplinen Rechtswissenschaft, Betriebs- oder Volkswirtschaft.

- Wenn er daran denkt, im Bereich *Forensischer Psychologie* zu arbeiten, müßte er gutachterliche Techniken in allen ihren Phasen erwerben und einüben, von der Klärung der Fragestellung bis zur Datenerhebung und ihrer Integration in einem Gutachtentext. Darüber hinaus sollte er seine Kenntnisse in Rechtspsychologie, in Kriminologie und in relevanten Ausschnitten der Rechtswissenschaft vertiefen.
- Wenn ihn die Aufgaben *Klinischer Psychologie* anlocken, sollte er seine diagnostischen Kenntnisse und Fertigkeiten ergänzen durch eine *therapeutische Ausbildung*.

Eine solche Ausbildung erweitert sein Tätigkeitsfeld und verbessert seine Aussicht auf einen Arbeitsplatz. Sie verlangt ihm aber auch Entscheidungen besonderer Art ab. Er kann sich einer der großen Therapieschulen anschließen, zum Beispiel der psychoanalytischen, der verhaltenstherapeutischen oder der sogenannten humanistischen Richtung (etwa der Gesprächs- oder der Gestalttherapie). Er kann es aber auch vorziehen, unterschiedliche therapeutische Modelle kennenzulernen, sie zu ‚prüfen‘ und sich erst ‚später‘ für eine bestimmte Therapieform zu entscheiden.

Was sollen diese Hinweise leisten? Sie sollen ermuntern zu einer Erweiterung des diagnostischen Grundwissens, das ein Lehrbuch vermittelt. Sie sollen Anregungen bieten für solche Studenten, die nach ihrem Studium in die psychologische Praxis gehen und sich dafür ein gerütteltes Maß diagnostischer Kenntnisse und Techniken erwerben wollen. Sie könnten auch dazu beitragen, den sogenannten ‚Praxischock‘ abzumildern, indem sie Wege weisen, die den Übergang vom Studium zur Praxis erleichtern.

Literatur

- Abels, D. (1965). Konzentrations-Verlaufs-Test (KVT) (2. Auflage). Göttingen: Hogrefe.
- ADAFI: Adaptiver Figurenfolgen-Lerntest: Guthke (1989).
- Aiken, L. R. (1982). Writing multiple choice items to measure higher-order educational objectives. *Journal of Educational and Psychological Measurement*, 42, 803-806.
- Algera, J. A. (1976). Reliability of selection interview data: A field study. *Nederlands Tijdschrift voor de Psychologie en haar Grensgebieden*, 31, 3-11.
- Althoff, K. (1984). Zur prognostischen Validität von Intelligenz- und Leistungstests im Rahmen der Eignungsdiagnostik. *Psychologie und Praxis*, 28, 144-148.
- Amelang, M. & Zielinski, W. (1994). *Psychologische Diagnostik und Intervention*. Berlin: Springer.
- American Psychological Association (APA). (1986). *Guidelines for computer-based tests and interpretations*. Washington: APA.
- American Psychological Association (APA). (1974). *Standards for educational and psychological tests (2nd Edition)*. Washington: APA.
- American Psychological Association (APA). (1992). *Ethical principles of psychologists and code of conduct*. *American Psychologist*, 47, 1597-1611.
- Armhauer, R. (1973). *Intelligenz-Struktur-Test 70 (IST 70)* (4. Auflage). Göttingen: Hogrefe.
- Andersen, E. B. (1982). Latent trait models and ability Parameter estimation. *Applied Psychological Measurement*, 6, 445-461.
- Anger, H. (1969). Befragung und Erhebung. In K. Gottschaldt, Ph. Lersch, F. Sander & H. Thomae (Hrsg.), *Handbuch der Psychologie in 12 Bänden*. (Bd. 7,1, herausgegeben von C.F. Graumann: Sozialpsychologie, S. 567-617). Göttingen: Hogrefe.
- Angleitner, A. (1976). *Methodische und theoretische Probleme bei Persönlichkeitsfragebogen mit einer ausführlichen Analyse deutschsprachiger Persönlichkeitsfragebögen*. Habilitationsschrift, Bonn.
- Angleitner, A. (1991). Personality psychology: Trends and developments. *European Journal of Personality*, 5, 185-197.
- Angleitner, A., Stumpf, H. & Wieck, Th. (1976). Die „Personality Research Form“ von Jackson: Konstruktion, bisheriger Forschungsstand und vorläufige Ergebnisse zur Äquivalenzprüfung einer deutschen Übersetzung. *Wehrpsychologische Untersuchungen*, 11, Heft 3.
- Angleitner, A. & Wiggins, J. S. (Eds.). (1986). *Personality assessment via questionnaire. Current issues in theory and measurement*. Berlin: Springer.
- APA: Siehe „American Psychological Association“.
- Arbeitskreis-AC: Siehe „Arbeitskreis Assessment Center“.
- Arbeitskreis Assessment Center. *Auswahl und Entwicklung von Führungskräften: Projektgruppe ‚Qualitätsstandards‘*. (Hrsg.). (1992). *Standards der Assessment-Center-Technik*. München. (Interessenten können die Standards gegen eine Bearbeitungsgebühr von 10,- DM beziehen bei Herrn Diplompsychologen Rainer Neubauer, Thierschstraße 23, 80538 München.)
- Arnold, W. (1972). Gutachten. In W. Arnold (Hrsg.), *Psychologisches Praktikum* (Bd. 2: Diagnostisches Praktikum, 7. Auflage, S. 322-327). Stuttgart: Fischer.

- Asch, S. E. (1946). Forming impressions of personality. *Journal of Abnormal and Social Psychology*, 41, 258-290.
- Atteslander, P. & Kneubühler, H.U. (1975). *Verzerrungen im Interview. Zu einer Fehlertheorie der Befragung*. Opladen: Westdeutscher Verlag.
- Atkinson, J. W. & McClelland, M.C. (1948). The projective expression of needs. II. The effect of different intensity of hunger drive on thematic apperception. *Journal of Experimental Psychology*, 38, 643-658.
- Axhausen, S. (1989). Projektive Verfahren. In E. Roth (Hrsg.) unter Mitarbeit von K. Heidenreich, *Sozialwissenschaftliche Methoden. Lehr- und Handbuch für Forschung und Praxis* (2., unwesentlich veränderte Auflage, S. 471-488). München: Oldenbourg.
- Bader, P., Hofmann, K. & Kubinger, K. D. (1993). Zur Brauchbarkeit der Normen von Papier-Bleistift-Tests für die Computer-Vorgabe: Ein Experiment am Beispiel des Gießen-Tests. *Zeitschrift für Differentielle und Diagnostische Psychologie*, 14, 129-135.
- Bagozzi, R. P. (1993). Assessing construct validity in personality research: Applications to measures of self-esteem. *Journal of Research in Personality*, 27, 49-87.
- Bales, R. F. (1950). *Interaction process analysis: A method for the study of small groups*. Cambridge, Mass.: Addison-Wesley.
- Bandura, A. (1973). *Aggression: A social learning analysis*. Englewood Cliffs: Prentice Hall.
- Bandura, A. (1977). *A social learning theory*. Englewood Cliffs: Prentice Hall. (Deutsch von H. Kober: (1979). *Sozial-kognitive Lerntheorie*. Stuttgart: Klett-Cotta).
- Bandura, A., Ross, D. & Ross, S.A. (1961). Transmission of aggression through imitation of aggressive models. *Journal of Abnormal and Social Psychology*, 63, 575-582.
- Barendregt, J. T. (1961). *Research in psychodiagnostics*. The Hague-Paris: Mouton.
- Bartel, H. (1971). *Statistik I*. - (1972). *Statistik II*. Stuttgart: Gustav Fischer.
- Bartenwerfer, H. (1983). Allgemeine Leistungsdiagnostik. In K. J. Groffmann & L. Michel (Hrsg.), *Enzyklopädie der Psychologie, Themenbereich B: Methodologie und Methoden, Serie II Psychologische Diagnostik* (Bd. 2: Intelligenz und Leistungsdiagnostik, S. 482-512). Göttingen: Hogrefe.
- Bastine, R. (1973). Zur Validität von Fragebögen. In G. Reinert (Hrsg.), *Bericht über den 27. Kongreß der Deutschen Gesellschaft für Psychologie in Kiel 1970*. (S. 63-67). Göttingen: Hogrefe.
- Baumann, U. (1990). Klassifikation. In U. Baumann & M. Perrez (Hrsg.), *Lehrbuch klinischer Psychologie*. Bd. 1: Grundlage, Diagnostik, Ätiologie (S. 50-55). Bern: Huber.
- Bäumler, G. (1974). *Lern- und Gedächtnistest (LGT 3)*. Göttingen: Hogrefe.
- Beaumont, J. G. & French, C. C. (1987). A clinical field study of eight automatic psychometric procedures: The Leichester-DHSS Project. *International Journal of Man-Machine-Studies*, 26, 661-681.
- Becker, H. & Langosch, I. (1990). *Produktivität und Menschlichkeit. Organisationsentwicklung und ihre Anwendung in der Praxis*. Stuttgart: Enke.
- Beckmann, D., Brähler, E. & Richter, H. E. (1990). *Der Gießen Test (GT). Ein Test für Individual- und Gruppendiagnostik*. Handbuch (4., überarbeitete Auflage mit Neustandardisierung). Bern: Huber.
- Behn, S. (1953). Über die Kunst des praktisch brauchbaren Gutachtens. *Psychologische Beiträge*, 1, 361-388.
- Behn-Eschenburg, H. (1952). Behn-Rorschach-Test (BERO-Test). Siehe: Zulliger, H. (1952). *Einführung in den Behn-Rorschach-Test* (3. Auflage). Bern: Huber.
- Bellak, L. & Bellak, S. S. (1949). The children's apperception test. New York: C. P. S. (Deutsch von W. Moog: (1955). *Kinder-Apperzeptions-Test (CAT)*. Göttingen: Hogrefe).
- Bellak, L. & Bellak, S. S. (1973). *The senior apperception technique*. New York: C. P. S.

- Benton, A. L. (1972). Benton-Test (4. Auflage). Bern: Huber. (Deutsche Bearbeitung von O. Spreen).
- Berkowitz, L. (1962). *Aggression: A social psychological analysis*. New York: McGraw-Hill.
- Berufsverband Deutscher Psychologen (BDP): Siehe „Deutscher Psychologen Verlag“.
- Binet, A. & Simon, T. (1905). *Methodes nouvelles pour le diagnostic du niveau intellectuel des anormaux*. *Année Psychologique*, 11, 191-241.
- Bierkens, P. (1968). *Die Urteilsbildung in der Psychodiagnostik*. München: Barth.
- Blau, G. (1962). Der psychologische Sachverständige im Strafprozeß, III. Das Gutachten. In G. Blau & E. Müller-Luckmann (Hrsg.), *Gerichtliche Psychologie* (S. 344-376). Berlin: Luchterhand.
- Blase, H. & Reeb, W. (1909). *Heinrichens Lateinisch-Deutsches Schulwörterbuch* (Achte Auflage neubearbeitet). Leipzig: Teubner.
- Bach, D. (1986). Das Gespräch mit dem Bewerber im Mittelpunkt gezielter Personalauswahl. *Psychologie und Praxis*, 30, 109-110.
- Bochenski, J. M. (1980). *Die zeitgenössischen Denkmethoden* (8. Auflage). München: Francke. (Uni-Taschenbücher: UTB 6).
- Boerner, K. (1993). *Das psychologische Gutachten*. Weinheim: Psychologie Verlags Union.
- Bohm, E. (1972). *Lehrbuch der Rorschach-Psychodiagnostik* (5. Auflage). Bern: Huber.
- Booth, J. F. (1995). Computerdiagnostik. In R. S. Jäger & F. Petermann (Hrsg.), *Psychologische Diagnostik* (3., korrigierte Auflage, S. 186-197). Weinheim: Beltz.
- Borkenau, P. & Ostendorf, F. (1993). NEO-Fünf-Faktoren Inventar (NEO-FFI) nach Costa und McCrae. Göttingen: Hogrefe.
- Bortz, J. (1984a). Befragen. In J. Bortz, *Lehrbuch der empirischen Forschung für Sozialwissenschaftler* (S. 163-189). Berlin: Springer.
- Bortz, J. (1984b). Beobachten. In J. Bortz, *Lehrbuch der empirischen Forschung für Sozialwissenschaftler* (S. 189-208). Berlin: Springer.
- Bortz, J. (1989). *Statistik für Sozialwissenschaftler* (3. Auflage). Berlin: Springer.
- Brem-Gräser, L. (1975). Familie in Tieren. Die Familiensituation im Spiegel der Kinderzeichnung. Entwicklung eines Testverfahrens (3. Auflage). München: Reinhardt.
- Brickenkamp, R. (Hrsg.). (1975). *Handbuch psychologischer und pädagogischer Tests*. Göttingen: Hogrefe.
- Brickenkamp, R. (Hrsg.). (1983). *Erster Ergänzungsband zum Handbuch psychologischer und pädagogischer Tests*. Göttingen: Hogrefe.
- Brickenkamp, R. (1994). Test d 2. Aufmerksamkeits-Belastungs-Test (8., erweiterte und neugestaltete Auflage). Göttingen: Hogrefe.
- Buggle, F. & Baumgärtel, F. (1972). *Hamburger Neurotizismus- und Extraversionsskala für Kinder und Jugendliche* (HANES, KJ). Göttingen: Hogrefe.
- Bühler, Ch. (1933). *Der menschliche Lebenslauf als psychologisches Problem*. Leipzig: Hirzel. (Neuaufgabe: 1959. Göttingen: Hogrefe).
- Bühler, Ch. (1969). Die allgemeine Struktur des menschlichen Lebenslaufs. In Ch. Bühler & F. Massarik (Hrsg.), *Lebenslauf und Lebensziele* (S. 10-22). Stuttgart: G. Fischer.
- Bühler, Ch. & Hetzer, H. (1972). *Kleinkindertest (BHKT)*. München: Barth.
- Bukasa, B., Kisser, R. & Wenninger, U. (1990). Computergestützte Leistungsdiagnostik bei verkehrspsychologischen Eignungsuntersuchungen. *Diagnostica*, 36, 148-165.
- Bundesanstalt für Arbeit (1991). *Berufswahltest (BWT)*, Handanweisung für die Berufsberatung. Nürnberg: Bundesanstalt für Arbeit.
- Bungard, W. unter Mitwirkung von R. H. Bay (1980). *Einführung in die psychologische Forschungspraxis. Kurseinheit 2: Beobachtung*. Hagen: Fernuniversität.
- Burgoon, M. & Ruffner, M. (1978). *Human communication*. New York: Holt, Rinehart & Winston.

- Buss, D. M. & Craig, K. H. (1983). The act frequency approach to personality. *Psychological Review*, 90, 105-126.
- Byrne, B. M. & Goffin, R. D. (1993). Modeling MTMM data from additive and multiplicative covariance structures: An audit of construct validity concordance. *Multivariate Behavioral Research*, 28, 67-96.
- Campbell, D. T. & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, 81, 81-105.
- Cantril, H. & Rugg, D. (1965). Die Formulierung von Fragen. In R. König (Hrsg.), *Das Interview* (S. 86-114). Köln: Kiepenheuer & Witsch.
- Cannell, C. F. & Fowler, F. J. (1967). Comparison of a self-enumerative procedure and a personal interview: A validity study. *Public Opinion Quarterly*, 27, 250-264.
- Cannell, C. F. & Kahn, R. L. (1968). Interviewing. In G. Lindzey & E. Aronson (Eds.), *Handbook of Social Psychology*. (Vol. 2, p. 526-595). Reading, Mass.: Addison-Wesley.
- Cary (1983). *SAS: Statistical Analysis System*. Institute Inc. Cary.
- Cattell, R. B. (1949). r_p and other coefficients of pattern similarity. *Psychometrika*, 14, 279-298.
- Cattell, R. B. (1958). What is 'objective' in objective personality tests? *Journal of Counseling Psychology*, 2, 285-289.
- Cattell, R. B. (1967). *The scientific analysis of personality* (2nd Edition). Harmondsworth: Penguin. (Deutsch von L. Piaggio: (1973). *Die empirische Erforschung der Persönlichkeit*. Weinheim: Beltz.)
- Cattell, R. B. (1969). The profile similarity coefficient r_p , in vocational guidance and diagnostic classification. *British Journal of Educational Psychology*, 39, 131-142.
- Cattell, R. B., Eber, H.B. & Tatsuka, M.M. (1970). *Handbook for the Sixteen Personality Factor Questionnaire* (16 PF). Ipat, Illinois.
- Cierpka, M. (1988). *Familiendiagnostik*. Berlin: Springer.
- Climont, C. E., Plutchik, R., Estrada, E., Gravia, L. F. & Arevalo, W. (1975). A comparison of traditional and symptom-checklist-based histories. *American Journal of Psychiatry*, 132, 450-453.
- Collins, M. & Odell, K. (1986). Computerization of a traditional test for nonverbal visual problem solving. *Cognitive Rehabilitation*, 4, 16-18.
- Conrad, W. (1983). Intelligenzdiagnostik. In K. J. Groffmann & L. Michel (Hrsg.), *Enzyklopädie der Psychologie, Themenbereich B: Methodologie und Methoden, Serie II: Psychologische Diagnostik* (Bd. 2: Intelligenz- und Leistungsdiagnostik, S. 104-201). Göttingen: Hogrefe.
- Cranach, M. von & Frenz, H. G. (1969). Systematische Beobachtung. In K. Gottschaldt, Ph. Lersch, F. Sander & H. Thomae (Hrsg.), *Handbuch der Psychologie in 12 Bänden* (Bd. 7,1, herausgegeben von C. F. Graumann: *Sozialpsychologie*, S. 269-331). Göttingen: Hogrefe.
- Cronbach, L.J. (1946). Response sets and test validity. *Educational and Psychological Measurement*, 6, 475-494.
- Cronbach, L. J. (1964). *Essentials of psychological testing*. New York: Harper & Row.
- Cronbach, L.J. & Gleser, G. C. (1965). *Psychological tests and personal decisions* (2nd Edition). Urbana: University of Illinois Press.
- Cronbach, L.J. & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, 52, 281-301.
- Cronbach, L.J., Gleser, G. C., Nanda, H. & Rajaratnam, N. (1972). The dependability of behavioral measurements: Theory of generalizability for scores and profiles. New York: John Wiley & Sons, Inc.
- Cronbach, L. J., Rajaratnam, N. & Gleser, G. C. (1963). Theory of generalizability: A liberalization of reliability theory. *British Journal of Statistical Psychology*, 16, 137-163.
- CSS: Complete Statistical System: StatSoft. Inc. 1986-1991.

- Dahl, G. (1971). Zur Berechnung des Schwierigkeitsindex bei quantitativ abgestufter Aufgabenbewertung. *Diagnostica*, 17, 139-142.
- Dahl, G. (1972). Reduzierter Wechsler-Intelligenztest (WIP). Meisenheim/Glan: Hain.
- Dahlstrom, W. G., Welsh, G. Sch. & Dahlstrom, L. (1972). *An MMPI Handbook* (2nd Edition). Minneapolis: University of Minnesota Press.
- Dahmer, H. & Dahmer, J. (1982). *Gesprächsführung. Eine praktische Einführung*. Stuttgart: Thieme.
- Daniels, J. C. (1971). *Figure Reasoning Test (FRT)* (6th. Edition). London: Crosby Lockwood & Son.
- Deegener, G. (1984). *Anamnese und Biographie im Kindes- und Jugendalter*. Weinheim: Beltz.
- Dehmelt, P., Kuhnelt, W. & Zinn, A. (1981). *Diagnostischer Elternfragebogen (DEF)* (4. Auflage). Weinheim: Beltz.
- Dailey, C. A. (1960). The life history as a criterion of assessment. *Journal of Counselling Psychology*, 7, 20-23.
- Deutscher Psychologen Verlag (DPV). (1986). *Berufsordnung für Psychologen*. Bonn: DPV.
- Deutscher Psychologen Verlag (DPV). (1994 a). *Leitsätze zur Dokumentation klinisch-psychologischer/psychotherapeutischer Interventionen*. Bonn: DPV.
- Deutscher Psychologen Verlag (DPV). (1994 b). *Richtlinien für die Erstellung Psychologischer Gutachten*. Bonn: DPV
- Dienel, P. C. (1978). *Die Planungszelle. Der Bürger plant seine Umwelt*. Opladen: Westdeutscher Verlag.
- Diesinger, I. (1977). *Der Affekttäter. Eine Analyse seiner Darstellung im forensisch-psychiatrischen Gutachten*. Berlin: de Gruyter.
- Dieterich, R. (1973). *Psychodiagnostik*. München: Reinhardt (Uni-Taschenbücher: UTB 273).
- Dilling, H., Mombour, W. & Schmidt, M. H. (Hrsg.). (1994). *ICD 10: „Internationale Klassifikation psychischer Störungen“ der Weltgesundheitsorganisation*. Bern: Huber.
- Dirks, H. (1961). Die Personalbeurteilung. In K. Gottschaldt, Ph. Lersch, F. Sander & H. Thoma (Hrsg.), *Handbuch der Psychologie in 12 Bänden* (Bd. 9 herausgegeben von A. Mayer & B. Herwig: *Betriebspsychologie*, S. 641-632). Göttingen: Hogrefe.
- Dollard, J., Doob, L., Miller, N. E., Mowrer, H. O. & Sears, R. R. (1939). *Frustration and Aggression*. New Haven: Yale University Press.
- Dörner, D. (1989). *Die Logik des Misslingens. Reinbek bei Hamburg*: Rowohlt.
- Dorsch, F. (1994). *Psychologisches Wörterbuch* (12., überarbeitete und erweiterte Auflage). (Herausgegeben von F. Dorsch †, H. Häcker & K.H. Stapf). Bern: Huber.
- DPV oder dpv: Siehe „Deutscher Psychologen Verlag“.
- Drey-Fuchs, Ch. (1958). *Der Fuchs-Rorschach-Test (FURO-Test)*. Göttingen: Hogrefe.
- Duhm, E. & Althaus, D. (1979). *Beobachtungsbogen für Kinder im Vorschulalter (BBK 4-6)*. Braunschweig: Westermann.
- Düker, H. & Lienert, G. A. (1965). *Konzentrations-Leistungs-Test (KLT)* (2. Auflage). Göttingen: Hogrefe.
- Durchholz, E. (1981). Der psychodiagnostische Prozeß. In E.G. Wehner (Hrsg.), *Psychodiagnostik in Theorie und Praxis* (S. 260-307). Frankfurt: Lang.
- Eberwein, M. (1992). *Verfahren der Klinischen Psychologie*. Herausgegeben von der Zentralstelle für Psychologische Information und Dokumentation. Trier: ZPID. Trier: Druckerei der Universität.
- Eckard, H.-H. (1977). Psychologische Diagnostik im Dienst beruflicher Beratung. In K. H. Seifert (Hrsg.), *Handbuch der Berufspsychologie* (S. 531-578). Göttingen: Hogrefe.
- Edwards, A. L. (1953). *Edwards personal preference Schedule*. New York: Psychological Corporation.

- Edwards, W. (1980). Multiattributive utility for evaluation. In M. W. Klein & K. S. Teilmann (Eds.), *Handbook of Criminal Justice Evaluation* (p. 177-215). Beverly Hills: Sage.
- Eggert, D. (1973). Probleme der Validität von Fragebogen. In G. Reinert (Hrsg.), *Bericht über den 27. Kongreß der Deutschen Gesellschaft für Psychologie in Kiel 1970* (S. 61-63). Göttingen: Hogrefe.
- Eggert, D. (1983).
- Eysenck-Persönlichkeits-Inventar. E-P-I. Handanweisung für die Durchführung und Auswertung (2., überarbeitete und ergänzte Auflage). Göttingen: Hogrefe.
- Eggert, D. (1994). *Lincoln-Oseretzký-Skala. Kurzform (LOS KF 18)* (2. Auflage). Weinheim: Beltz.
- Erdman, H.P., Klein, M. H., Greist, J. H., Bass, S.M., Bires, J. K. & Matchinger, P. E. (1987). A comparison of the diagnostic interview schedule and clinical diagnosis. *American Journal of Psychiatry*, 144, 1477-1480.
- Erven, H. (1981). *Mein Paradies. 32jährige Erfahrungen eines Praktikers im naturgemäßen Obst- und Gemüsebau*. Remagen: Erven.
- Esser, M. (1995). Selbsturteile. In W. Sarges (Hrsg.), *Management-Diagnostik* (2., vollständig überarbeitete und erweiterte Auflage, S. 802-809). Göttingen.
- Eysenck, H. J. (1953). Fragebogen als Meßmittel der Persönlichkeit. *Zeitschrift für experimentelle und angewandte Psychologie*, 1, 291-335.
- Eysenck, H. J. (1967). *Wege und Abwege der Psychologie*. Hamburg: Rowohlt. rde17.
- Eysenck, H. J. (1974). E-P-I. Vgl. Eggert.
- Eysenck, H. J. (1976). *Sex and personality*. London: Open Books. (Deutsch von L. Nürenberger: (1976). *Sexualität und Persönlichkeit*. Wien: Europa Verlag.)
- Eysenck, H. J. & Eysenck, M. W. (1985). *Personality structure and individual differences. A natural science approach* (2nd Revised Edition). London: Plenum Press. (Deutsch von H. D. Rosacker: (1987). *Persönlichkeit und Individualität. Ein naturwissenschaftliches Paradigma*. Mit einem Vorwort von Theo Herrmann. Weinheim: Psychologie Verlags Union.)
- Fahrenberg, J. (1964). Objektive Tests zur Messung der Persönlichkeit. In K. Gottschaldt, Ph. Lersch, F. Sander & H. Thomae (Hrsg.), *Handbuch der Psychologie in 12 Bänden* (Bd.6, herausgegeben von R. Heiß: *Psychologische Diagnostik*, S. 488-832). Göttingen: Hogrefe.
- Fahrenberg, J., Selg, H. & Hampel, R. (1978). *Freiburger Persönlichkeitsinventar (FPI): Handbuch* (3. Auflage). Göttingen: Hogrefe.
- Fahrenberg, J., Hampel, R. & Selg, H. (1989). *Freiburger Persönlichkeitsinventar (FPI und FPI-R): Handbuch* (5., ergänzte Auflage). Göttingen: Hogrefe.
- Faßnacht, G. (1979). *Systematische Verhaltensbeobachtung. Eine Einführung in die Methodologie und Praxis*. München: Reinhardt. (Uni-Taschenbücher: UTB 889).
- Faßnacht, G. (1995). *Systematische Verhaltensbeobachtung. Eine Einführung in die Methodologie und Praxis* (2., völlig neubearbeitete Auflage). München: Reinhardt.
- Fay, E. (1982). *Der „Test für medizinische Studiengänge“ (TMS). Ausgewählte Aspekte seiner Genese*. Braunschweig: Agentur Pedersen.
- Fay, E. (1993). *HAWIE-R. Hamburg-Wechsler-Intelligenztest für Erwachsene. Revision 1991*. Diagnostica, 39, 271-279.
- Feger, B. (1984). Die Generierung von Testitems zu Lehrtexten. Diagnostica, 30, 24-46.
- Fehnmann, U. (1995). Recht. In R. S. Jäger & F. Petermann (Hrsg.), *Psychologische Diagnostik* (3., korrigierte Auflage, S. 129-138). Weinheim: Beltz Psychologie Verlags Union.
- Fennekels, G. (1987). *Validität des Assessment-Centers bei Führungskräfteauswahl und -entwicklung*. Philosophische Dissertation, Bonn.
- Fischer, G. (1974). *Einführung in die Theorie psychologischer Tests*. Bern: Huber.
- Fisseni, H. J. (1974). *Zur Situation von Frauen in Altersheimen. Ergebnisse einer Tageslaufanalyse*. Philosophische Dissertation, Bonn.

- Fisseni, H. J. (1987). Ertragnisse biographischer Forschung in der Persönlichkeitspsychologie. In G. Jüttemann & H. Thomae (Hrsg.), *Biographie und Psychologie* (S. 149-265). Berlin: Springer.
- Fisseni, H. J. (1992). *Persönlichkeitsbeurteilung. Zu Theorie und Praxis des psychologischen Gutachtens* (2. Auflage). Mit einem Vorwort von Prof. Dr. H. Thomae. Göttingen: Hogrefe.
- Fisseni, H. J. (1995). Rückmeldung der Kandidatenbeurteilung an den Auftraggeber. In W. Sarges (Hrsg.), *Management-Diagnostik* (2., vollständig überarbeitete und erweiterte Auflage, S. 802-809). Göttingen: Hogrefe.
- Fisseni, H. J. & Fennekels, G. P. (1995). *Das Assessment-Center. Eine Einführung für Praktiker*. Göttingen: Hogrefe.
- Fisseni, H. J., Olbrich, E., Halsig, N., Mailahn, N. & Ittner, E. (1993). *Auswahlgespräche mit Medizinstudenten. Modelle - Erfahrungen - Vorschläge*. Göttingen: Hogrefe.
- Föderation Deutscher Psychologenvereinigungen: Siehe „Deutscher Psychologen Verlag“.
- Frances, A., Clarkin, J. F., Gilmore, M., Hurt, S. W. & Brown, R. (1984). Reliability of criteria for borderline personality disorder: A comparison of DSM-3 and the diagnostic interview for borderlines patients. *Journal of American Psychiatry*, 141, 1080-1083.
- Frank, L. K. (1939). Projective methods for the study of personality. *Journal of Psychology*, 8, 389-413.
- Frei, E. (1981). Psychologische Arbeitsanalyse. Eine Einführung zum Thema. In F. Frei & E. Ulich (Hrsg.), *Beiträge zur psychologischen Arbeitsanalyse* (S. 11-36). Bern: Huber.
- Freud, S. (1940). Abriß der Psychoanalyse. *Gesammelte Werke XVII* (S. 63-138). London: Imago.
- Fricke, R. (1973). Testgütekriterien bei lehrzielorientierten Tests. In P. Strittmatter (Hrsg.), *Lernzielorientierte Leistungsmessung* (S. 115-135). Weinheim: Beltz.
- Fricke, R. (1974). *Kriterienorientierte Leistungsmessung*. Stuttgart: Kohlhammer.
- Fricke, R. (1995). Videotests: „True-to-live“ - Testsituationen durch interaktives Video. In W. Sarges (Hrsg.), *Management-Diagnostik* (2., vollständig überarbeitete und erweiterte Auflage, S. 463-466). Göttingen: Hogrefe.
- Frieling, E. (1975). *Psychologische Arbeitsanalyse*. Stuttgart: Kohlhammer.
- Frieling, E. (1977). Die Arbeitsanalyse als Grundlage der Eignungsdiagnostik. In J.K. Triebe & E. Ulich (Hrsg.), *Beiträge zur Eignungsdiagnostik* (S. 20-90). Bern: Huber.
- Frieling, E. & Hoyos, C. Graf (1978). *Fragebogen zur Arbeitsanalyse (FAA)*. Bern: Huber.
- Frieling, E. & Sonntag, K. (1987). *Lehrbuch der Arbeitspsychologie*. Bern: Huber.
- Frinken, M. (1980). Die Bedeutung der Testgütekriterien für Interview und Exploration unter besonderer Berücksichtigung der Validität. Unveröffentlichte Psychologische Diplomarbeit, Bonn.
- Fuchs, Ch. (1958). Vgl. Drey-Fuchs, Ch. (1958). *Der Fuchs-Rorschach-Test. (FURO-Test)*. Göttingen: Hogrefe.
- Fuchs, W. (1982). *Biographische Forschung*. Hagen: Fernuniversität.
- Funke, J. (1993). Computergestützte Arbeitsproben: Begriffserklärung, Beispiele sowie Entwicklungspotentiale. *Zeitschrift für Arbeits- und Organisationspsychologie*, 3, 109-118.
- Funke, U. (1986). Die Validität verschiedener eignungsdiagnostischer Verfahren bei Lehrstellenbewerbern. *Psychologie und Praxis*, 30, 92-97.
- Funke, U. (1993). Computergestützte Eignungsdiagnostik mit komplexen dynamischen Szenarios. *Zeitschrift für Arbeits- und Organisationspsychologie*, 3, 119-129.
- Funke, U. & Schuler, H. (1986). Personalauswahl im Bereich industrieller Forschung und Entwicklung. *Psychologie und Praxis*, 30, 34-51.
- Fürntratt, E. (1969). *Differentieller Wissenstest*. Göttingen: Hogrefe.
- Gielen, D. & Kaden, S. (1977). Die Bedeutung der Testgütekriterien für die Exploration. Unveröffentlichte Diplomarbeit, Bonn.

- Giese, H. & Schmidt, G. (1968). Studentensexualität. Hamburg: Rowohlt.
- Gigerenzer, G. (1981). Messung und Modellbildung in der Psychologie. München: Reinhardt. (Uni-Taschenbücher: UTB 1047).
- Glaser, R. (1973). Ein kriteriumsbezogener Test. In P. Strittmatter (Hrsg.), Lernzielorientierte Leistungsmessung (S. 62-68). Weinheim: Beltz.
- Goldberg, L. (1971). A historical survey of personality scales and inventories. In P. McReynolds (Ed.), Advances in psychological assessment. (Vol. II, p. 293-336). Palo Alto, California: Science and behaviour books.
- Goldfried, M. R. & Kent, R. N. (1976). Herkömmliche gegenüber verhaltenstheoretischer Persönlichkeitsdiagnostik: ein Vergleich methodischer und theoretischer Voraussetzungen. In D. Schulte (Hrsg.), Diagnostik in der Verhaltenstherapie (2. Auflage, S. 3-23). München: Urban & Schwarzenberg.
- Göllner, D. & Deter, H.-C. W. (1980). Bemerkungen zu Verlaufs- und Erfolgskontrollen. In P. Hahn (Hrsg.), Die Psychologie des 20. Jahrhunderts (Bd. IX: Ergebnisse für die Medizin (1): Psychosomatik, S. 1003-1022). Zürich: Kindler.
- Gösslbauer, J. P. (1981). Grundprinzipien der Entscheidungstheorie in der psychologischen Diagnostik. In E.G. Wehner (Hrsg.), Psychodiagnostik in Theorie und Praxis (S. 214-257). Frankfurt: Lang.
- Graumann, C. F. (1967): vgl. Hörmann, H., Jäger, A. O. Kaminski, G., Cohen, R., Herrmann, Th. & Graumann, C. F. Die Beziehung zwischen psychologischer Diagnostik und Grundlagenforschung. Symposium III. In F. Merz (Hrsg.), Bericht über den 25. Kongreß der Deutschen Gesellschaft für Psychologie, Münster 1966 (S. 101-131). Göttingen: Hogrefe.
- Graumann, C. F. (1972). Interaktion und Kommunikation. In K. Gottschaldt, Ph. Lersch, F. Sander & H. Thomae (Hrsg.), Handbuch der Psychologie in 12 Bänden (Bd.7,2, herausgegeben von C. F. Graumann: Sozialpsychologie, S. 1109-1262). Göttingen: Hogrefe.
- Graumann, C. F. (1980). Psychologie - humanistisch oder human? In U. Völker (Hrsg.), Humanistische Psychologie (S. 39-51). Weinheim und Basel: Beltz.
- Grawe, K., Donati, R. & Bemauer, F. (1994). Psychotherapie im Wandel. Von der Konfession zur Profession. Göttingen: Hogrefe.
- Groffmann, K. J. (1982). Die Entwicklung der Intelligenzmessung. In K. J. Groffmann & L. Michel (Hrsg.), Enzyklopädie der Psychologie, Themenbereich B: Methodologie und Methoden, Serie II: Psychologische Diagnostik (Bd. 2: Intelligenz und Leistungsdiagnostik, S. 1-103). Göttingen: Hogrefe.
- Groffmann, K. J. & Michel, L. (Hrsg.). (1982a, 1982b, 1983a, 1983b). Enzyklopädie der Psychologie, Themenbereich B: Methodologie und Methoden, Serie II: Psychologische Diagnostik, Göttingen: Hogrefe.
- Bd. 1 (1982 a). Grundlagen psychologischer Diagnostik.
- Bd. 2 (1983 a). Intelligenz- und Leistungsdiagnostik.
- Bd. 3 (1982 b). Persönlichkeitsdiagnostik.
- Bd. 4 (1983 b). Verhaltensdiagnostik.
- Gross, A. (1985). Children. In M. Hersen & S.M. Turner (Eds.), Diagnostic interviewing (p. 309-335). New York: Plenum Press.
- Gross, L. D., Sallis, J. F., Buono, M. J. & Roby, J. J. (1990). Reliability of interviewers using the seven-day physical activity recall. Research Quarterly For Exercise And Sport, 61, 321-325.
- Grubitzsch, D. (1991). Testtheorie - Testpraxis. Psychologische Tests und Prüfverfahren im kritischen Überblick (vollständig überarbeitete und erweiterte Neuauflage). Hamburg: Rowohlt.rororo Sachbuch 8814.
- Guilford, J. P. (1936). The determination of item difficulty when chance success is a factor. Psychometrika, 1, 259-264.
- Guilford, J. P. (1946). New standards for test evaluation. Educational and Psychological Measurements, 6, 427-439.

- Guilford, J. P. (1959). *Personality*. New York: McGraw-Hill. (Deutsch von H. Kottenhoff & U. Agrell (1964). *Persönlichkeit. Logik, Methodik und Ergebnisse ihrer quantitativen Erforschung*. Weinheim: Beltz.)
- Guilford, J. P. & Hoepfner, R. (1971). *The analysis of intelligence*. New York: McGraw-Hill. (Deutsch von R. Horn: (1976). *Analyse der Intelligenz*. Weinheim: Beltz.)
- Guion, R. M. (1977). Content validity - the source of my discontent. *Applied Psychological Measurement*, 1, 1-10.
- Gulliksen, H. (1950). *Theory of mental tests*. New York: Wiley.
- Gunderson, E. K. E. & Kapfer, E. L. (1966). The predictive validity of clinical ratings for an extreme environment. *British Journal of Psychiatry*, 112, 405-441.
- Gunderson, J. G., Ronningstam, E. & Bodkin, A. (1990). The diagnostic interview for narcissistic patients. *Archives of General Psychiatry*, 47, 676-680.
- Guthke, J. (1972). *Zur Diagnostik der intellektuellen Lernfähigkeit*. Berlin: VEB.
- Guthke, J. (1989). *ADAFI (Adaptiver Figurenfolgen-Lerntest)*.
- Guthke, J., Böttcher, H. R. & Sprung, L. (Hrsg.). (1990, 1991). *Psychodiagnostik. Ein Lehr- und Arbeitsbuch für Psychologen sowie empirisch arbeitende Humanwissenschaftler*. Bd. 1 (1990). Bd. 2 (1991). Berlin: Deutscher Verlag der Wissenschaften.
- Guthke, J., Böttcher, H. R. & Sprung, L. (1991). *Psychodiagnostische Untersuchung-Begutachtung*. In J. Guthke, H. R. Böttcher, & L. Sprung, L. (Hrsg.), *Psychodiagnostik*. Bd. 2 (S. 246-336). Berlin: Deutscher Verlag der Wissenschaften.
- Guttman, L. (1944). A basis for scaling qualitative data. *American Sociological Review*, 9, 139-150.
- Haas, R. M. P. (1975). *Die Vermittlung inter- und intraindividuelle Persönlichkeitsunterschiede im Begutachtungsprozeß*. Philosophische Dissertation, Bonn.
- Häcker, H. (1982). Objektive Tests zur Messung der Persönlichkeit. In K. J. Groffmann & L. Michel (Hrsg.), *Enzyklopädie der Psychologie, Themenbereich B: Methodologie und Methoden, Serie II: Psychologische Diagnostik* (Bd. 3: *Persönlichkeitsdiagnostik*, S. 132-185). Göttingen: Hogrefe.
- Häcker, H. (1994 a). Fragebogen. In F. Dorsch, H. Häcker & K. H. Stapf (Hrsg.), *Psychologisches Wörterbuch* (S. 256). Bern: Huber.
- Häcker, H. (1994 b). Persönlichkeitsfragebogen. In F. Dorsch, H. Häcker & K. H. Stapf (Hrsg.), *Psychologisches Wörterbuch* (S. 560-561). Bern: Huber.
- Hageböck, J. (1994). *Computerunterstützte Diagnostik in der Psychologie*. Göttingen: Hogrefe.
- Hagen, Cornelia von (1988). *Formen der Auseinandersetzung mit chronisch dermatologischer Erkrankung*. Philosophische Dissertation, Bonn.
- Halder-Sinn, P. (1980). Effektivität psychotherapeutischer Intervention. In W. Wittling (Hrsg.), *Handbuch der Klinischen Psychologie* (Bd. 6: *Klinische Psychologie in Forschung und Praxis*, S. 92-115). Hamburg: Hoffmann und Kampe.
- Halsig, N. (1984). *Bewältigungsstrategien von Medizinstudenten in problematischen Situationen bei Studienbeginn*. Philosophische Dissertation, Bonn.
- Hank, G., Halweg, K. & Klann, N. (Hrsg.). (1990). *Diagnostische Verfahren für Berater*. Weinheim: Beltz.
- Hampel, R. & Selg, H. (1975). Fragebogen zur Erfassung von Aggressivitätsfaktoren (FAF). Göttingen: Hogrefe.
- Hartland, J. (1995). Sprache und Denken. In J. Gerstenmaier (Hrsg.), *Einführung in die kognitive Psychologie* (S. 195-245). München: Reinhardt.
- Hartmann, H. A. (1973). *Psychologische Diagnostik* (2., unveränderte Auflage). Stuttgart: Kohlhammer. Urban-Taschenbuch 135.
- Hartmann, H.A. (1984). Zur Ethik gutachterlichen Handelns. In H. A. Hartmann & R. Haubl (Hrsg.), *Psychologische Begutachtung* (S. 1-32). München: Urban & Schwarzenberg.

- Hartmann, H. A. & Haubl, R. (Hrsg.). (1984). *Psychologische Begutachtung*. München: Urban & Schwarzenberg.
- Hase, H. D. & Goldberg, L. R. (1967). Comparative validity of different strategies of constructing personality inventories scales. *Psychological Bulletin*, 67, 231-248.
- Hasemann, K. (1983). Verhaltensbeobachtung und Ratingverfahren. In K. J. Groffmann & L. Michel (Hrsg.), *Enzyklopädie der Psychologie, Themenbereich B: Methodologie und Methoden, Serie II: Psychologische Diagnostik* (Bd. 4: Verhaltensdiagnostik, S. 434-488). Göttingen: Hogrefe.
- Hathaway, S. R. & McKinley, J. C. (1951). *Minnesota Multiphasic Personality Inventory (MMPI) (Revised Edition)*. New York: Psychological Corporation.
- Hathaway, S. R. & McKinley, J. C. (1963). *Minnesota Multiphasic Personality Inventory (MMPI): MMPI Saarbrücken* (deutsche Bearbeitung: O. Spreen). Bern: Huber.
- Haubl, R. (1984). Praxeologische und epistemologische Aspekte psychologischer Begutachtung. In H. A. Hartmann & R. Haubl (Hrsg.), *Psychologische Begutachtung* (S. 33-74). München: Urban & Schwarzenberg.
- Haubl, R. & Spitznagel, A. (1983). Diagnostik sozialer Beziehungen. In K. J. Groffmann & L. Michel (Hrsg.), *Enzyklopädie der Psychologie, Themenbereich B: Methodologie und Methoden, Serie II: Psychologische Diagnostik* (Bd. 4: Verhaltensdiagnostik, S. 702-858). Göttingen: Hogrefe.
- Heidenreich, K. (1989). Die Verwendung standardisierter Test. In E. Roth (Hrsg.) unter Mitarbeit von K. Heidenreich, *Sozialwissenschaftliche Methoden. Lehr- und Handbuch für Forschung und Praxis* (2., unwesentlich veränderte Auflage, S. 400-416). München: Oldenbourg.
- Heil, F. E. (1995). Klinische versus statistische Urteilsbildung. In R. S. Jäger & F. Petermann (Hrsg.), *Psychologische Diagnostik. Ein Lehrbuch* (3., korrigierte Auflage, S. 39-42). Weinheim: Psychologie Verlags Union.
- Heinz, G. (1982). *Fehlerquellen forensischer psychiatrischer Gutachten*. Heidelberg: Kriminalistik Verlag.
- Heiß, R. (1964). Technik, Methodik und Problematik des Gutachtens. In K. Gottschaldt, Ph. Lersch, F. Sander & H. Thomae (Hrsg.), *Handbuch der Psychologie in 12 Bänden* (Bd. 6, herausgegeben von R. Heiß: *Psychologische Diagnostik*, S. 975-995). Göttingen: Hogrefe.
- Helzer, J. E., Robins, L. N., Taibleson, N., Woodruff, A. Jr., Reich, T. & Wish, E. D. (1977). Reliability of psychiatric diagnosis. *Archives of General Psychiatry*, 34, 129-133.
- Heneman, R. L. (1986). The relationship between supervisory ratings and results-oriented measures of performance: A metaanalysis. *Personnel Psychology*, 39, 811-826.
- Hermans, H., Petermann, F. & Zielinski, W. (1978). *Leistungs-Motivations-Test (LMT)*. Amsterdam: Swets & Zeitlinger.
- Hergovich, H. (1992). *Computer-Häuschentest*. Dissertation, Universität Wien, Wien.
- Hess, U. & Schmitt-Planert, A. (1985). Das Assessment-Center, ein eignungsdiagnostisches Hilfsmittel für die betriebliche Praxis. In W. Kugeman, S. Preiser & K. Schneewind (Hrsg.), *Psychologie und komplexe Lebenswirklichkeit. Festschrift zum 65. Geburtstag von Walter Toman* (S. 185-200). Göttingen: Hogrefe.
- Hilke, R. (1993). Computerunterstützte Eignungsdiagnostik im Psychologischen Dienst der Bundesanstalt für Arbeit. *Zeitschrift für Arbeits- und Organisationspsychologie*, 3, 138-141.
- Hiltmann, H. (1966). *Kompodium der psychodiagnostischen Tests* (2. Auflage). Bern: Huber.
- Hinrichs, J. R. (1978). An eight year follow-up of a management assessment center. *Journal of Applied Psychology*, 63, 569-601.
- Hodge, R. D., Andrews, D.A., Robinson, D. & Hollett, J. (1988). The construct validity of interview-based assessments in family counseling. *Journal of Clinical Psychology*, 44, 563-572.
- Hofer, P. J. & Green, B. F. (1985). The challenge of competence and creativity in computerized psychological testing. *Journal of Consulting and Clinical Psychology*, 53, 826-838.

- Holtzmann, W. H. (1961). Holtzmann Inkblot Technique: Guide to administration and scoring (2nd Edition). New York: The Psychological Corporation.
- Holtzmann, W. H. (1972). Holtzmann Inkblot Technik (HIT). (Deutsche Bearbeitung: Lehrbuch der HIT von H.A. Hartmann, L. von Rosenstiel & P. Neumann. Bern: Huber.)
- Honaker, L. M. (1988). The equivalence of computerized and conventional MMPI administration: A critical review. *Clinical Psychological Review*, 8, 561-577.
- Hörmann, H. (1982). Theoretische Grundlagen der projektiven Verfahren. In K. J. Groffmann & L. Michel (Hrsg.), *Enzyklopädie der Psychologie, Themenbereich B: Methodologie und Methoden, Serie II: Psychologische Diagnostik* (Bd. 1: Grundlagen der psychologischen Diagnostik, S. 173-247). Göttingen: Hogrefe.
- Horn, W. (1969). Prüfungssystem für Schul- und Bildungsberatung (PSB). Göttingen: Hogrefe.
- Horn, W. (1983). Leistungsprüfungssystem (LPS: 2. Auflage). Göttingen: Hogrefe.
- Hornby, A. S. (1989). (Chief Editor: A. P. Cowie). *Oxford advanced learners dictionary* (4th Edition). Oxford: Oxford University Press.
- Hornby, A. S., Gatenby, E. V. & Wakefield, H. (1960). *The advanced learners dictionary of current English* (12th Edition). London: University Press.
- Hornke, L. F. (1977). Antwortabhängige Testverfahren: Ein neuartiger Ansatz psychologischen Testens. *Diagnostica*, 23, 1-14.
- Hornke, L. F. & Rettig, K. (1989). Regelgeleitete Itemkonstruktion unter Zuhilfenahme kognitionspsychologischer Überlegungen. In K. D. Kubinger (Hrsg.), *Moderne Testtheorie - Ein Abriss samt neuesten Beiträgen* (2., verbesserte Auflage). Weinheim: Beltz.
- Hornthal, St. (1985 a). Das Bewerber-Interview. Praktische Hinweise zur Verbesserung des Interviews mittels Arbeitsproben und Tätigkeitssimulationen. *Personal*, 26-30.
- Hornthal, St. (1985 b). Das Bewerber-Interview mit Sachbearbeitern. *Personal*, 120-122.
- Horst, P. (1971). *Messung und Vorhersage*. Weinheim: Beltz.
- Hoyos, C. Graf (1986). Die Rolle der Anforderungsanalyse im eignungsdiagnostischen Prozeß. *Psychologie und Praxis*, 30, 59-67.
- Hron, A. (1982). Interview. In G. L. Huber & H. Mandl (Hrsg.), *Verbale Daten* (S. 119-140). Weinheim: Beltz.
- Huber, G.L. (1989). AQUAD. Auswertung qualitativer Daten, Version 2.2/1989. Manual zur computerunterstützten Auswertung qualitativer Daten mit dem Softwarepaket AQUAD / Version 2.2 (Bericht Nr. 25). Tübingen: Arbeitsbereich Pädagogische Psychologie, Institut für Erziehungswissenschaft, Universität Tübingen.
- Huber, G. L. & Mandl, H. (Hrsg.). (1982). *Verbale Daten. Eine Einführung in die Grundlagen und Methoden der Erhebung*. Weinheim: Beltz.
- Huber, H. P. (1973 a). *Psychometrische Einzelfalldiagnostik*. Weinheim: Beltz.
- Huber, H.P. (1973 b). Verallgemeinerungen des Cattellschen Profilähnlichkeitskoeffizienten r_p unter dem Aspekt der klassischen Reliabilitätstheorie. *Zeitschrift für experimentelle und angewandte Psychologie*, 20, 39-53.
- Huber, O. (1989). Beobachtung. In E. Roth (Hrsg.) unter Mitarbeit von K. Heidenreich, *Sozialwissenschaftliche Methoden. Lehr- und Handbuch für Forschung und Praxis* (2., unwesentlich veränderte Auflage, S. 124-143). München: Oldenbourg.
- Hubert, L.J. & Baker, F. B. (1978). Analyzing the multitrait-multimethod matrix. *Multivariate Behavioural Research*, 13, 163-179.
- Humboldt-Psychologie-Lexikon. (1990). Herausgegeben von der Redaktion Naturwissenschaft und Medizin des Bibliographischen Instituts. Mit einer Einleitung von Professor Dr. Peter R. Hofstätter. München: Humboldt-Taschenbuch-Verlag Jacobi.
- Hurt, S. W., Hyler, S. E., Frances, A., Clarkin, J. F. & Brent, R. (1986). Assessing borderline personality disorder with self-report, clinical interview, or semistructured interview. *Journal of American Psychiatry*, 141, 1228-1231.

- Irle, M. (1955). Berufs-Interessen-Test (BIT) (Neuaufgabe). Göttingen: Hogrefe.
- Irle, M. & Allehoff, W. (1984). Berufs-Interessen-Test II (BIT II) (2. Auflage). Göttingen: Hogrefe.
- Jackson, D. N. (1967). Personality Research Form. Manual. New York: Research Psychologists Press.
- Jackson, D.N. (1974). Manual for the Personality Research Form. Goshen: Research Psychologists Press.
- Jackson, D. N. (1975). The relative validity of scales prepared by naive writers and those based on empirical methods of personality scale construction. *Educational and Psychological Measurement*, 35, 361-370.
- Jackson, R. (1973). Die Entwicklung kriteriumsbezogener Tests. In P. Strittmatter (Hrsg.), *Lernzielorientierte Leistungsmessung* (S. 92-103). Weinheim: Beltz.
- Jäger, A. O. (1967). Dimensionen der Intelligenz. Göttingen: Hogrefe.
- Jäger, A. O. & Althoff, K. (1984). Der Wilde-Intelligenz-Test (WIT). Ein Strukturdiagnostikum. Göttingen: Hogrefe.
- Jäger, R. S. (1982). Diagnostische Urteilsbildung. In K. J. Groffmann & L. Michel (Hrsg.), *Enzyklopädie der Psychologie, Themenbereich B: Methodologie und Methoden, Serie II: Psychologische Diagnostik* (Bd. 1: Grundlagen psychologischer Diagnostik, S. 295-375). Göttingen: Hogrefe.
- Jäger, R. S. (1985). Diagnostik. In Th. Herrmann & E. D. Lantermann (Hrsg.), *Persönlichkeitspsychologie. Ein Handbuch in Schlüsselbegriffen* (S. 225-232). München: Urban & Schwarzenberg.
- Jäger, R. S. (1986). Der diagnostische Prozeß. Eine Diskussion psychologischer und methodischer Randbedingungen (2. Auflage). Göttingen: Hogrefe.
- Jäger, R. S. (1990). Computerdiagnostik - ein Überblick. *Diagnostica*, 36, 96-114.
- Jäger, R. S. (1992). Statusdiagnostik. In R. S. Jäger & F. Petermann (Hrsg.), *Psychologische Diagnostik* (3., korrigierte Auflage, S. 200-202). Weinheim: Psychologie Verlags Union.
- Jäger, R. S. & Krieger, W. (1994). Zukunftsperspektiven der computerunterstützten Diagnostik, dargestellt am Beispiel der treatmentorientierten Diagnostik. *Diagnostica*, 40, 217-243.
- Jäger, R.S., Lischer, S., Münster, B. & Ritz, B. unter Mitarbeit von H. Fuchs-Entzminger (1976). *Biographisches Inventar zur Diagnose von Verhaltensstörungen (BIV)*. Göttingen: Hogrefe.
- Jäger, R. S. & Petermann, F. (Hrsg.). (1995). *Psychologische Diagnostik* (3., korrigierte Auflage). Weinheim: Psychologie Verlags Union.
- Jäger, R. S. & Scheurer, H. (1995). Prozeßdiagnostik. In R.S. Jäger & F. Petermann (Hrsg.), *Psychologische Diagnostik* (3., korrigierte Auflage, S. 202-208). Weinheim: Psychologie Verlags Union.
- Jahoda, M., Deutsch, M. & Cook, St. W. (1965). Die Technik der Auswertung: Analyse und Interpretation. In R. König (Hrsg.), *Das Interview* (S.271-289). Köln: Kiepenheuer & Witsch.
- Janke, W. (1973a). Über Konstruktion von Fragebogen. In G. Reinert (Hrsg.), *Bericht über den 27. Kongreß der Deutschen Gesellschaft für Psychologie in Kiel 1970* (S.41-44). Göttingen: Hogrefe.
- Janke, W. (1973 b). Das Dilemma von Persönlichkeitsfragebogen. In G. Reinert (Hrsg.), *Bericht über den 27. Kongreß der Deutschen Gesellschaft für Psychologie in Kiel 1970* (S.44-48). Göttingen: Hogrefe.
- Janke, W. (1982). Klassenzuordnung: Zuordnung von Personen zu vorgegebenen Klassen. In K. J. Groffmann & L. Michel (Hrsg.), *Enzyklopädie der Psychologie, Themenbereich B: Methodologie und Methoden, Serie II: Psychologische Diagnostik* (Bd. 1: Grundlagen psychologischer Diagnostik, S. 376-466). Göttingen: Hogrefe.

- Jeserich, W. (1990). Mitarbeiter auswählen und fördern. Assessment-Center-Verfahren (5., unveränderter Nachdruck). München: Hanser.
- Jessnitzer, K. (1988). Der gerichtliche Sachverständige (9. Auflage). Köln: Heymann.
- Jochmann, W. (1984). Der implizite diagnostische Prozeß in der Personalberatung und seine aussagenlogische Formalisierung. *Psychologie und Praxis*, 28, 119-129.
- Jochmann, W. (1987). Einzelbeurteilungsprogramm - Ein Beispiel für die Absicherung von Einstellungs- und Beförderungsentscheidungen. *Kienbaum informiert*. 2/87, 1-3.
- Joerger, K. (1973). Gruppentest für die soziale Einstellung (SET) (2. Auflage). Göttingen: Hogrefe.
- Jüttemann, G. & Thomae, H. (Hrsg.). (1987). Biographie und Psychologie. Berlin: Springer.
- Kaegi, A. (1904). Benselers Griechisch-Deutsches Schulwörterbuch (zwölfte, erweiterte und vielfach verbesserte Auflage). Leipzig: Teubner.
- Kalinowsky-Czech, M. (1984). Selbst- und Fremdauswertung von Explorationen zur Erfassung des Selbstbildes. Philosophische Dissertation, Bonn.
- Kaminski, G. (1970). Verhaltenstheorie und Verhaltensmodifikation. Entwurf einer integrativen Theorie psychologischer Praxis am Individuum. Stuttgart: Klett.
- Kaminski, G. (1977). Beobachtung. In Th. Herrmann, P. R. Hofstätter, H. P. Huber & F. E. Weinert (Hrsg.), *Handbuch psychologischer Grundbegriffe* (S. 68-73). München: Kösel.
- Kamp, L.J. Th. van der (1976). Generability and educational measurement. In D.N.M. de Gruijter & L.J.Th. van der Kamp (Eds.), *Advances in psychological and educational measurement* (p. 173-184). London, New York: Wiley.
- Kasubek, W. & Aschenbrenner, K. M. (1978). Optimierung subjektiver Urteile: Anwendung der multiattributiven Nutzentheorie bei medizinischen Therapieentscheidungen. *Zeitschrift für experimentelle und angewandte Psychologie*, 25, 594-616.
- Keil, W. (1973). Reaktionseinstellung und Fragebogenkonstruktion. In G. Reinert (Hrsg.), *Bericht über den 27. Kongreß der Deutschen Gesellschaft für Psychologie in Kiel 1970* (S. 53-58). Göttingen: Hogrefe.
- Kelly, G.A. (1955). The psychology of personal constructs. Vol. I & II. New York: Norton.
- Kelly, G. A. (1958). Man's construction of his alternatives. In G. Lindzey (Ed.), *The assessment of human motives* (p. 33-64). New York: Rinehart.
- Kemmler, L. (1974). Die Anamnese in der Erziehungsberatung (3. Auflage). Bern: Huber.
- Keßler, B. H. (1982). Biographische Diagnostik. In K. J. Groffmann & L. Michel (Hrsg.), *Enzyklopädie der Psychologie, Themenbereich B: Methodologie und Methoden, Serie II: Psychologische Diagnostik*. (Bd. 3: Persönlichkeitsdiagnostik, S. 1-56). Göttingen: Hogrefe.
- Kipnowski, A. (1981). Hinweise auf die Gestaltung psychologischer Gutachten. *Psychologie und Praxis*, 4, 183-190.
- Kirusek, T. J. & Sherman, R. E. (1968). Goal attainment scale: a general method for evaluating comprehensive community mental health program. *Community Mental Health Journal*, 4, 443-453.
- Kisker, K. P. (1969). Phänomenologie der Intersubjektivität. In K. Gottschaldt, Ph. Lersch, F. Sander & H. Thomae (Hrsg.), *Handbuch der Psychologie in 12 Bänden* (Band 7.1, herausgegeben von C. F. Graumann: Sozialpsychologie, S. 81-107). Göttingen: Hogrefe.
- Kisser, R. (1986). Computertestgeräte. In R. Brickenkamp (Hrsg.), *Handbuch der apparativen Verfahren in der Psychologie* (S. 469-515). Göttingen: Hogrefe.
- Kisser, R. (1995). Adaptive Strategien. In R. S. Jäger & F. Petermann (Hrsg.), *Psychologische Diagnostik* (3., korrigierte Auflage, S. 161-170). Weinheim: Psychologie Verlags Union.
- Klapprott, J. (1975). Einführung in die psychologische Methodik. Stuttgart: Kohlhammer. Urban-Taschenbücher 214.
- Klauer, K. J. (1978). Perspektiven der pädagogischen Diagnostik. In K. J. Klauer (Hrsg.), *Handbuch der pädagogischen Diagnostik*. (Bd. 1, S. 3-14). Düsseldorf: Schwann.

- Klauer, K. J. (1984). Kontentvalidität. *Diagnostica*, 30, 1-23.
- Klauer, K. J. (1987). Kriteriumsorientierte Tests. *Lehrbuch der Theorie und Praxis lehrzielorientierten Messens*. Göttingen: Hogrefe.
- Klein, J. (1982). Die Rechtmäßigkeit psychologischer Tests im Personalbereich. Gelsenkirchen: Mannhold.
- Kleinevoss, R. (1978). Zum Kommunikationsprozeß zwischen Diagnostiker und Auftraggeber. Philosophische Dissertation, Berlin.
- Klieme, E. & Stumpf, H. (1990). Computereinsatz in der pädagogisch-psychologischen Diagnostik. In K. Ingenkamp & R. S. Jäger (Hrsg.), *Tests und Trends* 8 (S. 13-63). Weinheim: Beltz.
- Klopfer, B. & Davidson, H. H. (1942). *The Rorschach Technique. An introductory manual*. New York: Harcourt, Brace & World. (Deutsch von H. Huber: (1974). *Das Rorschach-Verfahren. Eine Einführung* (2. Auflage). Bern: Huber.)
- Kluwe, R. H. (1995). Computergestützte Systemsimulationen. In W. Sarges (Hrsg.), *Management-Diagnostik* (2., vollständig überarbeitete und erweiterte Auflage, S. 458-463). Göttingen: Hogrefe.
- Koch, K. (1972). *Der Baum-Test* (6. Auflage). Bern: Huber.
- Kohli, M. (1978). ‚Offenes‘ und ‚geschlossenes‘ Interview: Neue Argumente zu einer alten Kontroverse. *Soziale Welt*, 29, 1-25.
- Kolko, D. J. & Kazdin, A. E. (1989). Assessment of dimensions of childhood firesetting among patients and nonpatients: The firesetting risk interview. *Journal of Abnormal Child Psychology*, 17, 157-177.
- Kompa, A. (1984). *Personalbeschaffung und Personalauswahl*. Stuttgart: Enke.
- König, R. (Hrsg.) unter Mitarbeit von D. Rüschemeyer & E. K. Scheuch (1965). *Das Interview. Formen, Technik, Auswertung*. Köln: Kiepenheuer & Witsch.
- Kornadt, H. J. & Zumkley, H. (1982). Thematische Apperzeptionsverfahren. In K. J. Groffmann & L. Michel (Hrsg.), *Enzyklopädie der Psychologie, Themenbereich B: Methodologie und Methoden. Serie II: Psychologische Diagnostik* (Bd. 3: Persönlichkeitsdiagnostik, S. 258-372). Göttingen: Hogrefe.
- Kötter, S. & Nordmann, E. (1987). Die Beobachtungsmethoden. In M. Cierpka (Hrsg.), *Familiendiagnostik* (133-152). Berlin: Springer.
- Krampen, G. (1981). *IPC-Fragebogen zu Kontrollüberzeugungen*. Göttingen: Hogrefe.
- Kranz, H. T. (1981). *Einführung in die klassische Testtheorie* (2. Auflage). Frankfurt: Fachbuchhandlung für Psychologie. Verlagsabteilung.
- Kratzmeier, H. unter Mitarbeit von R. Horn (1978). *Raven-Matrizen-Test. Standard Progressive Matrices (SPM). Deutsche Bearbeitung*. Weinheim: Beltz.
- Kristof, W. (1957). Zur Frage der statistischen Sicherung von Profildifferenzen. *Zeitschrift für experimentelle und angewandte Psychologie*, 4, 692-696.
- Kriz, J. (1991). *Grundkonzepte der Psychotherapie* (3. Auflage). Ein Einführung. Weinheim: Psychologie Verlags Union.
- Kruse, A. (1987). Biographische Methode und Exploration. In G. Jüttemann & H. Thomae (Hrsg.), *Biographie und Psychologie* (S. 119-137). Berlin: Springer.
- Kubinger, K. D. (1987). *Adaptives Testen*. In R. Horn, K. Ingenkamp & R. S. Jäger (Hrsg.), *Tests und Trends* 6 (S. 103-127). München: Psychologie Verlags Union.
- Kubinger, K. D. (Hrsg.). (1989). *Moderne Testtheorie. Ein Abriß samt neuesten Beiträgen* (2., verbesserte Auflage). Weinheim: Beltz.
- Kubinger, K. D. (1993). Testtheoretische Probleme der Computerdiagnostik. *Zeitschrift für Arbeits- und Organisationspsychologie*, 3, 130-137.
- Kubinger, K. D. (1995 a). Testtheorie: Probabilistische Modelle. In R. S. Jäger & F. Petermann (Hrsg.), *Psychologische Diagnostik* (3., korrigierte Auflage, S. 322-334). Weinheim: Beltz.

- Kubinger, K. D. (1995 b). Einführung in die Psychologische Diagnostik. Weinheim: Psychologie Verlags Union.
- Kubinger, K. D. (1995c). Gutachtenerstellung. In K. D. Kubinger, Einführung in die Psychologische Diagnostik (S. 261-282). Weinheim: Beltz.
- Kubinger, K. D. & Farkas, M. G. (1991). Die Brauchbarkeit der Normen von Papier-Bleistift-Tests für die Computer-Vorgabe: Ein Experiment am Beispiel der SPM von Raven als kritischer Beitrag. Zeitschrift für Differentielle und Diagnostische Psychologie, 12, 257-266.
- Kubinger, K. D. & Wurst, E. (1994). Adaptives Intelligenz-Diagnosticum (AID) (2., überarbeitete Auflage). Weinheim: Beltz.
- Kuliga, K. (1990). Quotex - Fragebogensystem. Simbach: ZAK.
- Küpper, Th. (1993). Gütekriterien von Explorationen in der Literatur des letzten Dezenniums. Unveröffentlichte Psychologische Diplomarbeit, Bonn.
- Kurth, W. (1980). Das Gutachten. Anleitung für Mediziner, Psychologen, Juristen. München: Reinhardt.
- Lane, J. W., Pollard, C. A. & Cox, G. L. (1990). Validity study of the anxiety symptoms interview. Journal of Clinical Psychology, 46, 52-57.
- Landy, F. J. (1976). The validity of the interview in police officer selection. Journal of Applied Psychology, 61, 193-198.
- Langner, R. (1989). DSM-III-X (Experten- und Lehrsystem zur Psychiatrischen Diagnostik auf der Grundlage des DSM-III-R). Weinheim: Beltz Test Gesellschaft.
- Laplanche, J. & Pontalis, J. B. (1967). Vocabulaire de la Psychoanalyse. Paris: Presses Universitaires de France. (Deutsch von E. Moersch: (1977). Das Vokabular der Psychoanalyse (3. Auflage). Frankfurt: Suhrkamp Taschenbuch Verlag. 2 Bände.)
- Lechler, P. (1982). Kommunikative Validierung. In G. L. Huber & H. Mandl (Hrsg.), Verbale Daten (S. 243-258). Weinheim: Beltz.
- Leeb, B., Hahlweg, K., Goldstein, M. J., Feinstein, E., Mueller, U., Dose, M. & Magana-Amato, A. (1991). Cross-national reliability, concurrent validity, and stability of a brief method for assessing expressed emotion. Psychiatry Research, 39, 25-31.
- Lehr, U. (1964). Diagnostische Erfahrungen aus explorativen Untersuchungen bei Erwachsenen. Psychologische Rundschau, 14, 97-106.
- Lehr, U., Thomae, H. & Schmitz-Scherzer, R. (1972). Psychologischer Befund, subjektiver Gesundheitszustand, internistischer Befund. Ärztliche Praxis, 90, 4393-4401.
- Lehrenkrauss, E. (1986). Der richtige Einsatz von Arbeitsproben bei der Personalauswahl. Personal, 281-283.
- Leichner, R. (1975). Zur Verarbeitung psychiatrischer Informationen. Diagnostica, 21, 147-166.
- Leichner, R. (1976). Zur Verarbeitung psychiatrischer Informationen. Diagnostica, 22, 163-180.
- Leichner, R. (1979). Psychologische Diagnostik. Grundlagen, Kontroversen, Praxisprobleme. Weinheim: Beltz.
- Leichner, R. (1983). Diagnostik. In R. Asanger & G. Wenninger (Hrsg.), Handwörterbuch der Psychologie (S. 80-85) (3. Auflage). Weinheim: Beltz.
- Lennep, D.J. van (1948). Four picture test. Den Haag: Martinus Nijhoff.
- Lennep, D.J. van & Houwink, R. H. (1958). Manual four picture test (2nd Edition). Utrecht: Nederlandse Stichting voor Psychotechniek.
- Lennertz, E. (1973). Thesen zur Itemsammlung bei Persönlichkeitsfragebogen. In G. Reinert (Hrsg.), Bericht über den 27. Kongreß der Deutschen Gesellschaft für Psychologie in Kiel 1970 (S. 58-61). Göttingen: Hogrefe.
- Lewin, K. (1951). Field theory in social sciences (edited by D. Cartwright). New York: Harper. (Deutsch: (1963). Feldtheorie in den Sozialwissenschaften. Ausgewählte theoretische Schriften. Bern: Huber.)

- Liebe], H. & Uslar, W. v. (1975). *Forensische Psychologie. Eine Einführung*. Stuttgart: Kohlhammer. Urban-Taschenbücher 219.
- Lienert, G. A. (1969). *Testaufbau und Testanalyse* (3. Auflage). Weinheim: Beltz.
- Lienert, G. A. & Raatz, U. (1994). *Testaufbau und Testanalyse* (5., völlig neubearbeitete und erweiterte Auflage). Weinheim: Beltz.
- Lindner, K. (1980). Die Überprüfbarkeit des Konkordanzmaßes „Ü“. *Zeitschrift für Empirische Pädagogik*, 4, 45-58.
- Lindzey, G. (1967). Thematic Apperception Test: Interpretative assumptions and related empirical evidence. In D.N. Jackson & S. Messick (Eds.), *Problems in human assessment*. p. 575-593). New York: McGraw-Hill.
- Links, P. S., Steiner, M., Offord, D.R. & Eppel, A. (1985). Stability of the diagnostic interview for borderlines diagnosis. *American Journal of Psychiatry*, 142, 1525.
- Linstone, H. A. & Turoff, M. (Eds.). (1975). *The Delphi method*. London: Addison-Wesley.
- Lippert, S. & Zeidler, M. (1986). Interviewsystem für Führungskräfte. *Personal*, 7, 284-288.
- Loevinger, J. (1948). The technique of homogeneous tests compared with some aspects of scale analysis and factor analysis. *Psychological Bulletin*, 45, 507-530.
- Lohaus, A. (1989). Datenerhebung in der Entwicklungspsychologie. Problemstellungen und Forschungsperspektiven. Bern: Huber.
- Lord, F. M. & Novick, M. R. (1974). *Statistical theories of mental test scores* (2nd Edition). Reading, Mass.: Addison-Wesley.
- Lorenz, J. H. (1987). *Lernschwierigkeiten und Einzelfallhilfe. Schritte im diagnostischen und therapeutischen Prozeß*. Göttingen: Hogrefe.
- Loretto, V. (1986). Effective interviewing is based on more than intuition. *Personnel Journal*, 65, 101-107.
- Lück, H. E. & Timaeus, E. (1969). Skalen zur Messung Manifester Angst (MAS) und sozialer Wünschbarkeit (SDS-E und SDS-CM). *Diagnostica*, 15, 134-141.
- Lückert, H.-R. (1964). Persönlichkeitsgutachten. In H.R. Lückert (Hrsg.), *Handbuch der Erziehungsberatung*. (Bd. 1, S. 361-418). München: Reinhardt.
- Lutz, R. (1978). *Das verhaltensdiagnostische Interview*. Stuttgart: Kohlhammer. Urban-Taschenbuch 262.
- Lutz, R. & Windheuser, H. J. (1976) Therapiebegleitende Diagnostik. In D. Schulte (Hrsg.), *Diagnostik in der Verhaltenstherapie* (2. Auflage, S. 196-218). München: Urban & Schwarzenberg.
- Maccoby, E. E. & Maccoby, N. (1965). Das Interview: Ein Werkzeug der Sozialforschung. In R. König (Hrsg.), *Das Interview. Formen, Technik, Auswertung* (4.Auflage, S.37-85). Köln: Kiepenheuer & Witsch.
- Magnusson, D. (1969). *Testtheorie*. Wien: Deuticke.
- Maier, O. (1980). Zur Informationspflicht bei psychologischen Eignungsuntersuchungen. *Psychologie und Praxis*, 24, 49-57.
- Marco, G. L. (1981). Equating tests in an era of test disclosure. In B. F. Green (Ed.), *Issues in testing: Coaching, disclosure, and ethic bias* (p. 105-122). San Francisco: Jossey-Bass.
- Mash, E. J. & Terdal, L. T. (1976). (Eds). *Behavior Therapie Assessment*, New York: Springer. (Deutsch von W. Stifter: (1980). *Kompodium der verhaltenstherapeutischen Diagnostik*. Frankfurt: Fachbuchhandlung für Psychologie.)
- Mason, M. A. & Belt, J. A. (1986). Effectiveness of specificity in recruitment advertising. *Journal of Management*, 12, 425-432.
- Matarazzo, J. D. (1983). Computerized psychological testing. *Science*, 221, Editorial.
- Maukisch, H. (1986). Erfolgskontrollen von Assessment Center-Systemen: Der Stand der Forschung. *Psychologie und Praxis*, 30, 86-91.

- McBride, J. R. & Martin, J.T. (1983). Reliability and validity of adaptive ability tests in a military setting. In D.J. Weiss (Ed.), *New horizons in testing* (p.224-236). New York: Academic Press.
- McDonald, D. A., Nussbaum, D. S. & Bagby, R. M. (1991). Reliability, validity and utility of the fitness interview test. *The Canadian Journal of Psychiatry*, 36, 480-484.
- Meehl, P. E. (1954). *Clinical versus statistical prediction. A theoretical analysis and a review of the evidence*. Minneapolis: University of Minnesota Press.
- Meehl, P. E. (1967). What can the clinician do well? In D.N. Jackson & S. Messick (Eds.), *Problems in human assessment* (p. 594-599). New York: McGraw-Hill.
- Mees, U. & Selg, H. (Hrsg.). (1977). *Verhaltensbeobachtung und Verhaltensmodifikation*. Stuttgart: Klett.
- Meili, R. (1961). *Lehrbuch der psychologischen Diagnostik* (4. Auflage). Bern: Huber.
- Meinefeld, W. (1977). *Einstellung und soziales Handeln*. Reinbek: Rowohlt.
- Meinefeld, W. (1983). Einstellung. In R. Asanger & G. Wenninger (Hrsg.), *Handwörterbuch der Psychologie* (3. Auflage, S. 92-99). Weinheim: Beltz.
- Meyerhoff, H. & Dony, M. (1970). Die Zuverlässigkeit anamnestischer Angaben zur frühkindlichen Entwicklung. *Zeitschrift für Kinderheilkunde*, 108, 41-45.
- Michel, L. & Conrad, W. (1982). Theoretische Grundlagen psychometrischer Tests. In K. J. Groffmann & L. Michel (Hrsg.), *Enzyklopädie der Psychologie, Themenbereich B: Methodologie und Methoden, Serie II: Psychologische Diagnostik* (Band 1: Grundlagen psychologischer Diagnostik, S. 1-129). Göttingen: Hogrefe.
- Michel, L. & Iseler, A. (1968). Beziehungen zwischen klinischen und psychometrischen Methoden der diagnostischen Urteilsbildung. In K. J. Groffmann & K. H. Wewetzer (Hrsg.), *Person als Prozeß* (S. 115-156). Bern: Huber.
- Michel, L. & Mai, N. (1968). Entscheidungstheorie und Probleme der Diagnostik bei Cronbach und Gleser. *Diagnostica*, 14, 99-120.
- Mischel, W. (1981). *Introduction to personality* (3rd Edition). New York: Holt, Rinehart & Winston.
- Mischel, W. (1993). *Introduction to personality* (5th Edition). New York: Harcourt Brace.
- Mittenecker, E. (1982). Subjektive Tests zur Messung der Persönlichkeit. In K. J. Groffmann & L. Michel (Hrsg.), *Enzyklopädie der Psychologie, Themenbereich B: Methodologie und Methoden, Serie IJ: Psychologische Diagnostik* (Bd. 3: Persönlichkeitsdiagnostik, S. 57-131). Göttingen: Hogrefe.
- Moser, K. (1987). Inhaltsvalidität als Kriterium psychologischer Tests. *Diagnostica*, 33, 110-122.
- Müller, A. (1973). *Verkehrs-Verständnis-Test (VVT)*. Homburg/Saar: Selbstverlag.
- Müller, G. F. & Nachreiner, F. (1988). Zur Anwendung der multi-atributiven Nutzentechnik bei der Personalauswahl. *Zeitschrift für Personalforschung*, 119-129.
- Mummendey, H. D. (1987). *Die Fragebogen-Methode. Grundlagen und Anwendung in Persönlichkeits-, Einstellungs- und Selbstkonzeptforschung*. Göttingen: Hogrefe.
- Murray, H.A. (1938). *Explorations in personality*. New York: Oxford University Press.
- Murray, H. A. (1943). *Thematic Apperception Test Manual*. Cambridge: Harvard University Press.
- Murray, H.A. & Morgan, C. D. (1935). A method for investigating phantasies: The Thematic Apperception Test. *Archives of Neurology and Psychiatry*, 34, 289-306.
- Murstein, B. I. (1963). *Theory and research in projective techniques (emphasizing the TAT)*. New York: Wiley.
- Murstein, B. I. (1965). *Handbook of projective techniques*. New York: Basic Books.
- Murstein, B. I. & Pryer, R. S. (1959). The concept of projection: A review. *Psychological Bulletin*, 56, 353-374.

- Neubauer, A.C., Urban, E. & Malle, B. F. (1991). Ravens Advanced Progressive Matrices: Computerunterstützte Präsentation versus Standardvorgabe. *Diagnostica*, 37, 204-212.
- Noelle, E. (1968). Umfragen in der Massengesellschaft. Einführung in die Methoden der Demoskopie. Hamburg: Rowohlt. rde 177/178.
- Noack, H. & Petermann, F. (1995). Entscheidungstheorie. In R. S. Jäger & F. Petermann (Hrsg.), *Psychologische Diagnostik* (3., korrigierte Auflage, S. 286-310). Weinheim: Beltz Psychologie Verlags Union.
- Norusis, M. J. (1986). (I) SPSS/PC+. - (II) Advanced statistics. - (III) Tables. Chicago, Illinois: SPSSinc.
- Nowotny, B., Schlote-Sauter, B. & Rey, E. R. (1990). Entwicklung eines strukturierten Angstinterviews für Senioren. *Zeitschrift für Gerontologie*, 23, 218-225.
- Nußbaum, A. (1987). Das Modell der Generalisierbarkeitstheorie. In K. J. Klauer, *Kriteriumsorientierte Test* (S. 114-136). Göttingen: Hogrefe.
- Osgood, C. E. & Suci, G. (1952). A measure of relation determined by both mean difference and profile information. *Psychological Bulletin*, 49, 251-262.
- Ostendorf, F., Angleitner, A. & Ruch, W. (1986). Die Multitrait-Multimethod Analyse. Göttingen. Hogrefe.
- Parry, H. J. & Crossley, H. M. (1950). Validity of responses to survey questions. *Public Opinion Quarterly*, 1, 61-80.
- Pawlik, K. (1976). Modell- und Praxisdimensionen psychologischer Diagnostik. In K. Pawlik (Hrsg.), *Diagnose der Diagnostik* (S. 13-43). Stuttgart: Klett.
- Pelzmann, S. (1972). Experimente zum Einfluß des psychologischen Befundes auf das psychiatrische Urteil. *Philosophische Dissertation*, Graz.
- Perrez, M. & Baumann, U. (Hrsg.). (1991). *Lehrbuch Klinische Psychologie*. Bd. 2: Intervention. Bern: Huber.
- Perrez, M. & Baumann, U. (1991). Systematik der klinisch-psychologischen Intervention: Einleitung. In M. Perrez & U. Baumann, (Hrsg.), *Lehrbuch Klinische Psychologie*. Bd. 2: Intervention (S. 21-30). Bern: Huber.
- Pervin, L. A. (1975). *Personality: Theory, assessment and research* (2nd Edition). New York: Wiley. (Deutsch von G. Schäfer-Kilius & H. Kilius: (1981). *Persönlichkeitstheorien*. München: Reinhardt.)
- Petermann, F. (1978). *Veränderungsmessung*. Stuttgart: Kohlhammer.
- Petermann, F. (1985). *Psychologie des Vertrauens*. Salzburg: Otto Müller.
- Petermann, F. (1986). Probleme und neuere Entwicklungen der Veränderungsmessung - ein Überblick. *Diagnostica*, 32, 4-16.
- Petermann, F. (1989). Die Messung von Veränderung. In E. Roth (Hrsg.) unter Mitarbeit von K. Heidenreich, *Sozialwissenschaftliche Methoden. Lehr- und Handbuch für Forschung und Praxis* (2., unwesentlich veränderte Auflage, S. 583-594). München: Oldenbourg.
- Petermann, F. (1995). Kontrollierte Praxis. In R. S. Jäger & F. Petermann (Hrsg.), *Psychologische Diagnostik* (3., korrigierte Auflage, S. 147-154). Weinheim: Beltz Psychologie Verlags Union.
- Petermann, F. & U. (1978). *Training mit aggressiven Kindern*. München: Urban & Schwarzenberg.
- Petermann, F. & U. (1980). *Erfassungsbogen für aggressives Verhalten in konkreten Situationen (EAS)*. Braunschweig: Westermann.
- Petermann, F. & U. (1987). *Training mit Jugendlichen. Förderung von Arbeits- und Sozialverhalten*. München: Psychologie Verlags Union.
- Pfäfflin, F. (1978). *Vorurteilsstruktur und Ideologie psychiatrischer Gutachten über Sexualstraftäter*. Stuttgart: Enke.
- Phillipson, H. (1955). *The object relations technique*. London: Tavistock Publications LTD.

- Pulver, U., Lang, A. & Schmid, F. W. (Hrsg.). (1978). *Ist Diagnostik Verantwortbar?* Bern: Huber.
- Quitmann, H. (1991). *Humanistische Psychologie* (2. Auflage). Zentrale Konzepte und philosophischer Hintergrund. Göttingen: Hogrefe.
- Rasch, G. (1960). Probabilistic models for some intelligence and attainment tests. Copenhagen: Danmarks Paedagogiske Institut.
- Rauch, M., Weber, W. & Wildgrube, W. (1993). Computergestützte Testdiagnostik im Psychologischen Dienst der Bundeswehr. *Zeitschrift für Arbeits- und Organisationspsychologie*, 3, 142-145.
- Rauchfleisch, U. (1979). *Handbuch zum Rosenzweig-Picture-Frustration-Test (PFT)*. - Bd. 1: Grundlagen, bisherige Resultate und Anwendungsmöglichkeiten des PFT. - Bd. 2: Manual zur Durchführung des PFT und Neueichung der Testformen für Kinder und Erwachsene. Bern: Huber.
- Rauchfleisch, U. (1991). *Kinderpsychologische Tests*. Stuttgart: Enke.
- Raven, J. C. (1971). *Standard Progressive Matrices* (13th Edition). London: Lewis. (Deutsche Bearbeitung von H. Kratzmeier unter Mitarbeit von R. Horn: (1978). *Raven-Matrizen-Test*. *Standard Progressive Matrices (SPM)*. Weinheim: Beltz.)
- Reckase, M. D. (1974). An interactive computer program for tailored testing based on the one-parameter logistic model. *Behavior Research Methods and Instrumentation*, 6, 208-212.
- Rehberg, J. (Hrsg.). (1976). *Probleme des gerichtspsychiatrischen und -psychologischen Gutachtens*. Diessenhofen: Rüegger.
- Reibnitz, U. von. (1983). Die Szenario-Technik. In H. Haase & K. Köppler (Hrsg.), *Fortschritte in der Marktpsychologie*. (Bd. 3: Grundlagen - Methoden - Anwendung, S. 111-133). Frankfurt: Fachbuchhandlung für Psychologie.
- Reinecker, H. (1991). Verhaltenstherapeutisch orientierte Intervention. In M. Perez & U. Baumann (Hrsg.), *Lehrbuch Klinische Psychologie*. Bd. 2: Intervention (S. 129-145). Bern: Huber.
- Rennen-Allhoff, B. (1991). Wie verlässlich sind Elternangaben? *Praxis Kinderpsychologie Kinderpsychiatrie*, 40, 333-338.
- Rettig, K. & Hornke, L. F. (1995). *Adaptives Testen*. In W. Sarges (Hrsg.), *Management-Diagnostik* (2., vollständig überarbeitete und erweiterte Auflage, S. 444-450). Göttingen: Hogrefe.
- Revers, W. (1958). *Der Thematische Apperzeptionstest (TAT)*. *Handbuch zur Verwendung des TAT in der psychologischen Persönlichkeitsdiagnostik*. Bern: Huber.
- Revers, W. (unter Mitarbeit von F. Popp & K. Täuber). (1979). *Der Thematische Apperzeptionstest (TAT)*. *Handbuch zur Verwendung des TAT in der psychologischen Persönlichkeitsdiagnostik* (4. [gegenüber der 3. von 1973], unveränderte Auflage). Bern: Huber.
- Revers, W. & Allesch, Ch. G. (1985). *Handbuch zum Thematischen Gestaltungstest (Salzburg)*. Weinheim: Beltz.
- Revers, W. & Widauer, H. (1985). *Thematischer Gestaltungstest (Salzburg)*. Weinheim: Beltz Test Gesellschaft.
- Ringelband, J. O. & Birkhan, G. (1995). Rückmeldung der Eignungsbeurteilung an den Kandidaten und diskursive Abstimmung. In W. Sarges (Hrsg.), *Management-Diagnostik* (2., vollständig überarbeitete und erweiterte Auflage, S. 796-802). Göttingen: Hogrefe.
- Rock, D. L. & Nolen, P. A. (1982). Comparison of the standard and computerized version of the Raven Coloured Progressive Matrices Test. *Perceptual and Motor Skills*, 54, 40-42.
- Roest, F. & Horn, R. (1990). *Mailbox-90: Computergestützte Diagnostik im Assessment Center*. *Diagnostica*, 36, 213-219.
- Roidt, G. H. & Haladyna, T. (1982). *A technology for test-item writing*. New York: Academic Press.

- Rollett, B. & Bartram, M. (1977). Anstrengungsvermeidungstest (AVT). Braunschweig: Westermann.
- Rorer, L. G. (1965). The great response-style myth. *Psychological Bulletin*, 63, 129-156.
- Rorschach, H. (1972). *Psychodiagnostik. Methodik und Ergebnisse eines wahrnehmungsdiagnostischen Experimentes*. (Deutemassen von Zufallsformen) (9. Auflage). Bern: Huber.
- Rosenzweig, S. (1945). The picture-association method and its application in a study of reactions to frustration. *Journal of Personality*, 14, 23-23. (Deutsche Bearbeitung: (A) Hörmann, H. & Moog, W. (1957). Form für Erwachsene. Göttingen: Hogrefe. - (B) Duhm, E. & Hansen, J. (1957). Form für Kinder. Göttingen: Hogrefe.)
- Rosenzweig, S., Fleming, E. E. & Rosenzweig, L. (1948). The children's form of the picture frustration study. *Journal of Psychology*, 26, 141-191.
- Roth, E. (Hrsg.) unter Mitarbeit von K. Heidenreich (1989). *Sozialwissenschaftliche Methoden. Lehr- und Handbuch für Forschung und Praxis* (2., unwesentlich veränderte Auflage). München: Oldenbourg.
- Rotter, J. B. & Hochreich, D.J. (1975). *Personality*. Glenview, Ill.: Scott, Foresman. (Deutsch von P. Baumann-Frankenberger: (1979). *Persönlichkeit. Theorien, Messung, Forschung*. Berlin: Springer).
- Rutter, M. & Graham, P. (1968). The reliability and validity of the psychiatric assessment of the child. I: Interview with the child. *British Journal of Psychiatry*, 114, 579-663.
- Rütter, T. (1978). Formen der Testaufgabe. In K. J. Klauer (Hrsg.), *Handbuch der Pädagogischen Diagnostik*. Bd. 1 (S. 257-280). Düsseldorf: Schwann.
- Sacher, J. & Fletcher, J. D. (1978). Administering paper-and-pencil tests by computer, or the medium is not always the message. In D.J. Weiss (Ed.), *Proceedings of the 1977 computerized adaptive testing conference* (p. 403-420). Minneapolis: University of Minnesota, Department of Psychology.
- Sarges, W. (1995). Interviews. In W. Sarges (Hrsg.), *Management-Diagnostik* (2., vollständig überarbeitete und erweiterte Auflage, S. 471-489). Göttingen: Hogrefe.
- SAS: Statistical Analysis System: Institute Inc. Cary 1983.
- Sauermann, P. (1979). *Betriebspsychologie*. Stuttgart: Enke.
- Sawyer, J. (1966). Measurement and prediction, clinical and statistical. *Psychological Bulletin*, 66, 178-200.
- Schaller, S. & Schmidtke, A. (1983). Verhaltensdiagnostik. In K. J. Groffmann & L. Michel (Hrsg.), *Enzyklopädie der Psychologie, Themenbereich B: Methodologie und Methoden, Serie II: Psychologische Diagnostik* (Bd. 4: Verhaltensdiagnostik, S. 489-701). Göttingen: Hogrefe.
- Scheiblechner, H. (1971). A simple algorithm for CML-parameter-estimation in Rasch probabilistic measurement model with two or more categories of answers. *Wien: Research Bulletin 5 des Psychologischen Instituts der Universität*.
- Scherer, K. R. & Walbott, H. G. (Hrsg.). (1984) *Nonverbale Kommunikation* (2. Auflage). Basel: Beltz.
- Scheuch, E. K. (1967). Das Interview in der Sozialforschung. In R. König (Hrsg.), *Handbuch der empirischen Sozialforschung* (Bd. 1, 2. Auflage, S. 136-196). Stuttgart: Enke.
- Scheuch, E.K. (1973). Das Interview in der Sozialforschung. In R. König (Hrsg.), *Handbuch der empirischen Sozialforschung* (Bd. 2, 3. Auflage, S. 66-190). Stuttgart: Enke.
- Schmale, H. & Schmidtke, H. (1966). *Berufseignungstest (BET)*. Bern: Huber.
- Schmalt, H.-D. (1976). *Leistungsmotivations-Gitter (LM-Gitter)*. Göttingen: Hogrefe.
- Schmid, F. W. (1995). Ethik. In R. S. Jäger & F. Petermann (Hrsg.), *Psychologische Diagnostik* (3., korrigierte Auflage, S. 121-128). Weinheim: Beltz Psychologie Verlags Union.
- Schmid, F. W. (1995). Einzel-Assessment. In W. Sarges (Hrsg.), *Management-Diagnostik* (2., vollständig überarbeitete und erweiterte Auflage, S. 703-716). Göttingen: Hogrefe.

- Schmidt, K. J. (1980). Die Bedeutung der Testgütekriterien für Interview und Exploration unter besonderer Berücksichtigung der Objektivität und Reliabilität. Unveröffentlichte Psychologische Diplomarbeit, Bonn.
- Schmidt, L. R. (1982). Diagnostische Begutachtung. In K. J. Groffmann & L. Michel (Hrsg.), *Enzyklopädie der Psychologie, Themenbereich B: Methodologie und Methoden, Serie II: Psychologische Diagnostik* (Bd. 1: Grundlagen psychologischer Diagnostik, S. 467-537). Göttingen: Hogrefe.
- Schmidt, L. R. (1995). Psychodiagnostisches Gutachten. In R. S. Jäger & F. Petermann (Hrsg.), *Psychologische Diagnostik* (3., korrigierte Auflage, S. 468-478). München-Weinheim: Psychologie Verlags Union.
- Schmidt, L. R. & Keßler, B. H. (1976). Anamnese: Methodische Probleme, Erhebungsstrategien und Schemata. Weinheim: Beltz.
- Schmidtchen, St. (1975). Psychologische Tests für Kinder und Jugendliche. Göttingen: Hogrefe.
- Schmitt, N., Coyle, B.W. & Saari, B.B. (1977). A review and critique of analyses of multi-trait-multimethod matrices. *Multivariate Behavioral Research*, 12, 447-478.
- Schmitt, N. & Stults, D.M. (1986). Methodology review: Analysis of multitrait-multimethod matrices. *Applied Psychological Measurement*, 10, 1-22.
- Schneewind, K.A. (1969). *Methodisches Denken in der Psychologie*. Bern: Huber.
- Schneewind, K.A., Schröder, G. & Cattell, R. B. (1983). *Der 16-Persönlichkeits-Faktoren-Test (16 PF)*. Bern: Huber.
- Schober, S. (1977). Einschätzung und Anwendung projektiver Verfahren in der heutigen klinisch-psychologischen Praxis. *Diagnostica*, 23, 364-372.
- Scholz, O. B. (1978). *Diagnostik in Ehe- und Partnerschaftskrisen*. München: Urban & Schwarzenberg.
- Scholz, O. B. (1987). *Ehe- und Partnerschaftsstörungen*. Stuttgart: Kohlhammer.
- Schoppe, K. J. (1975). *Verbaler Kreativitätstest (VKT)*. Göttingen: Hogrefe.
- SPSS: Statistical Package for the Social Science: Siehe Norusis.
- Schraml, W. (1964). Das Psychodiagnostische Gespräch (Exploration und Anamnese). In K. Gottschaldt, Ph. Lersch, F. Sander & H. Thomae (Hrsg.), *Handbuch der Psychologie in 12 Bänden* (Bd. 6 herausgegeben von R. Heiß: *Psychologische Diagnostik*, S. 868-897). Göttingen: Hogrefe.
- Schröder, R.-D. (1976). Das diagnostische Urteil. Eine wissenschaftstheoretische Analyse und eine Untersuchung relevanter Forschungsansätze. Philosophische Dissertation, Mannheim.
- Schuler, H. & Stehle, W. (1987). Assessment-Center als Methode der Personalentwicklung. *Beiträge zur Organisationspsychologie* (Bd. 3). Göttingen: Hogrefe.
- Schulte, D. (1976). Der diagnostisch-therapeutische Prozeß in der Verhaltenstherapie. In D. Schulte (Hrsg.), *Diagnostik in der Verhaltenstherapie* (2. Auflage, S. 60-73). München: Urban & Schwarzenberg.
- Schwarzer, R. (1981). *Streß, Angst und Hilflosigkeit*. Stuttgart: Kohlhammer. (Darin: *Hilflosigkeits- und Selbstwirksamkeitsskala*.)
- Schwenkmezger, P. & Hank, P. (1993). Papier-Bleistift- versus computerunterstützte Darbietung von State-Trait-Fragebogen: eine Äquivalenzüberprüfung. *Diagnostica*, 39, 189-210.
- Seek, IJ. (1982). Die Aussagequalität der Exploration nach empirischen Studien. Unveröffentlichte Psychologische Diplomarbeit, Bonn.
- Sehringer, W. (1982). Zeichnerische und spielerische Gestaltungsverfahren. In K. J. Groffmann & L. Michel (Hrsg.), *Enzyklopädie der Psychologie, Themenbereich B: Methodologie und Methoden, Serie II: Psychologische Diagnostik* (Bd. 3: *Persönlichkeitsdiagnostik*, S. 430-528). Göttingen: Hogrefe.
- Seidenstücker, E. & G. (1974). Interviewforschung: Allgemeiner Teil. In W. J. Schraml & U. Baumann (Hrsg.), *Klinische Psychologie* (Bd. 2: *Methoden, Ergebnisse und Probleme*, S. 377-402). Bern: Huber.

- Seiffert, H. (1971 a). Einführung in die Wissenschaftstheorie 1 (4. Auflage). Sprachanalyse - Deduktion - Induktion in Natur- und Sozialwissenschaften. München: C. H. Beck. Beck'sche Schwarze Reihe, Bd. 60.
- Seiffert, H. (1971 b). Einführung in die Wissenschaftstheorie 2 (3., unveränderte Auflage). Geisteswissenschaftliche Methoden: Phänomenologie - Hermeneutik und historische Methode - Dialektik. München: C. H. Beck. Beck'sche Schwarze Reihe, Bd. 61.
- Seitz, W. (1977). Persönlichkeitsbeurteilung durch Fragebogen. Braunschweig: Westermann.
- Seitz, W. & Rausche, A. (1976). Persönlichkeitsfragebogen für Kinder zwischen 9 und 14 Jahren (PFK 9-14). Braunschweig: Westermann.
- Selg, H. & Bauer, W. (1971). Forschungsmethoden der Psychologie. Stuttgart: Kohlhammer. Urban-Taschenbücher 121.
- Sines, L. K. (1959). The relative contribution of four kinds of data to accuracy in personality assessment. *Journal of Consulting Psychology*, 23, 483-492.
- Six, B. (1975). Die Relation von Einstellung und Verhalten. *Zeitschrift für Sozialpsychologie*, 6, 270-296.
- Sixtl, F. (1967). Meßmethoden der Psychologie. Theoretische Grundlagen und Probleme. Weinheim: Beltz.
- Skinner, B. F. (1948). *Walden Two*. New York: Macmillan.
- Skinner, B. F. (1953). *Science and human behavior*. New York: Macmillan.
- Skre, I., Onstad, S., Torgersen, S. & Kringlen, E. (1991). High interrater reliability for the structured clinical interview for DSM-III-R Axis I (SDIC-I). *Acta Psychiatrica Scandinavica*, 84, 167-173.
- Sloane, R. B., Staples, F. R., Cristol, A. H., Yorkston, N.J. & Wipple, K. (1975 a). *Psychotherapy versus behavior therapy*. Cambridge: Harvard University Press.
- Sloane, R. B., Staples, F. R., Cristol, A. H., Yorkston, N.J. & Wipple, K. (1975 b). Short-term analytically oriented psychotherapy versus behavior therapy. *American Journal of Psychiatry*, 132, 373-377.
- Solomon, J. & Starr, B. (1968): *School Apperception Method (SAM): Manual*. New York: Springer.
- Sommer, G. (1973). Die Problematik der Erfassung von ‚Konzentration‘, dargestellt am KLT (Konzentrations-Leistungs-Test). *Diagnostica*, 19, 62-75.
- Sonnenberg, H.-G. (1993). Computergestützte psychologische Diagnoseverfahren bei der Auswahl von Führungskräften. *Zeitschrift für Arbeits- und Organisationspsychologie*, 3, 146-149.
- Soskin, W. F. (1959). Bias in postdiction from projective tests. *Journal of Abnormal Psychology*, 58, 69-78.
- Spearman, C. (1927). *The abilities of man*. London: Macmillan.
- Spearman, C. (1938). Measurement of intelligence. *Scientifica*, 64, 75-82.
- Spitznagel, A. (1964). Die diagnostische Situation. Ein Beitrag zur Theorie und Psychologie der Datengewinnung. Unveröffentlichte Habilitationsschrift. Freiburg.
- Spitznagel, A. (1982 a). Die diagnostische Situation. In K. J. Groffmann & L. Michel (Hrsg.), *Enzyklopädie der Psychologie, Themenbereich B: Methodologie und Methoden, Serie II: Psychologische Diagnostik*. (Band 1: Grundlagen psychologischer Diagnostik, S. 248-294). Göttingen: Hogrefe.
- Spitznagel, A. (1982 b). Grundlagen, Ergebnisse und Probleme der Formdeutungsverfahren. In K. J. Groffmann & L. Michel (Hrsg.), *Enzyklopädie der Psychologie, Themenbereich B: Methodologie und Methoden, Serie II: Psychologische Diagnostik* (Band. 3: Persönlichkeitsdiagnostik, S. 186-257). Göttingen: Hogrefe.
- Spitznagel, A. (1984). Kommunikationspsychologische Forschungsergebnisse zur Produktion und Rezeption von Gutachtentexten. In H. A. Hartmann & R. Haubl (Hrsg.), *Psychologische Begutachtung* (S. 127-159). München: Urban & Schwarzenberg.

- Spitznagel, A. (1990). Projektive Verfahren. In W. Sarges (Hrsg.), *Management-Diagnostik* (2., vollständig überarbeitete und erweiterte Auflage, S. 515-525). Göttingen: Hogrefe.
- Spörli, S. (1978). *Kritische Theorie diagnostischer Praxis - dargestellt am Beispiel Verkehrspsychologie*. Bern: Huber.
- Staabs, G. von (1964). *Der Scenotest* (3. Auflage). Bern: Huber.
- Staples, F. R., Sloane, R. B., Cristol, A. H., Yorkston, N.J. & Wipple, K. (1975). Differences between behavior therapists and psychotherapists. *Archives of General Psychiatry*, 32, 1517-1522.
- Staples, F. R., Sloane, R. B., Whipple, K., Cristol, A. H. & Yorkston, N.J. (1976). Process and outcome in psychotherapy and behavior therapy. *Journal of Consulting and Clinical Psychology*, 44, 340-350.
- Stauffer, E. & Trottmann-Gschwend, A. (1980). *Geist-Bilder-Interessen-Inventar (GBII)*. Lisse: Swets & Zeitlinger.
- Steck, P. (1989). Thematischer Gestaltungstest. (Rezension). *Diagnostica*, 35, 276-282.
- Steck, P. (1991). Bemerkungen zu L. Tents Beitrag „Psychodiagnostische Verfahren und die minima scientifica“. *Diagnostica*, 37, 89-92.
- Steck, P. (1993). Gutachten. In A. Schorr (Hrsg.), *Handwörterbuch der Angewandten Psychologie* (S. 320-323). Bonn: Deutscher Psychologen Verlag.
- Steinberg, M., Rounsaville, B. & Cicchetti, D. V. (1990). The structured clinical interview for DSM-III-R dissociative disorders: Preliminary report on a new diagnostic instrument. *American Journal of Psychiatry*, 147:1, 76-82.
- Stelz, I. (1972). Was bringt das Rasch-Modell für die Praxis? *Psychologische Beiträge*, 14, 298-310.
- Stern, W. (Hrsg.). (1904). *Beiträge zur Psychologie der Aussagen*. Leipzig: Barth.
- Stern, W. (1921). *Die Differentielle Psychologie in ihren methodischen Grundlagen* (3. Auflage). Leipzig: Barth.
- Stern, W. (1923). *Die menschliche Persönlichkeit* (3. Auflage). Leipzig: Barth.
- Stern, W. (1926). *Jugendliche Zeugen in Sittlichkeitsprozessen*. Leipzig: Quelle & Meyer.
- Stevens, S. S. (1963). Mathematics, measurement, and psychophysics. In S. S. Stevens (Ed.), *Handbook of Psychology* (5th Edition, p. 1-49). New York: Wiley.
- Stoll, F. (1977). Zur Abhängigkeit des Eignungsdiagnostikers und des Probanden: Lösungsvorschläge. In J. K. Triebe & E. Ulich (Hrsg.), *Beiträge zur Eignungsdiagnostik* (S. 203-214). Bern: Huber.
- Stumpf, H., Angleitner, A., Wieck, Th., Jackson, D.N. & Belloch-Till, H. (1985). *Deutsche Personality Research Form (PRF)*. Göttingen: Hogrefe.
- Süllwold, F. & Berg, M. (1967). *Problemfragebogen für Jugendliche*. Göttingen: Hogrefe.
- Tanzer, N. (1987). Möglichkeiten und Probleme des Einsatzes von Personal-Computern in der klinisch-psychologischen Begutachtung. *Schlußbericht des 27. Kongresses des Bundesverbandes österreichischer Psychologen* (B. P.). Wien: AUVA.
- Taylor, H. C. & Russell, J. T. (1939). The relationship of validity coefficients to the practical effectiveness of tests in selection: Discussion and tables. *Journal of Applied Psychology*, 23, 565-578.
- Tent, L. (1991). Psychodiagnostische Verfahren und die minima scientifica. *Diagnostica*, 37, 83-88.
- Testkuratorium der Föderation Deutscher Psychologenvereinigungen (1986). Beschreibung der einzelnen Kriterien für die Testbeurteilung. *Diagnostica*, 32, 358-360.
- Testkuratorium der Föderation Deutscher Psychologenvereinigungen (1986). Richtlinien für den Einsatz elektronischer Datenverarbeitung in der psychologischen Diagnostik. *Psychologische Rundschau*, 37, 163-165.

- Tewes, U. (Hrsg.). (1991). Hamburg-Wechsler-Intelligenztest für Erwachsene. Revision 1991. (HAWIE-R). Bern: Huber.
- Thiel, R., Keller, G. & Binder, A. (1979). Arbeitsverhaltensinventar (AVI). Braunschweig: Westermann.
- Thomae, H. (1967). Prinzipien und Formen der Gestaltung psychologischer Gutachten. In K. Gottschaldt, Ph. Lersch, F. Sander & H. Thomae (Hrsg.), Handbuch der Psychologie in 12 Bänden (Bd. 11, herausgegeben von U. Undeutsch: Forensische Psychologie, S. 743-767). Göttingen: Hogrefe
- Thomae, H. (1968). Das Individuum und seine Welt. Göttingen: Hogrefe.
- Thomae, H. (1977). Psychologie in der modernen Gesellschaft. Hamburg: Hoffmann & Campe.
- Thomae, H. (1987). Psychologische Biographik als Synthese idiographischer und nomothetischer Forschung. In G. Jüttemann und H. Thomae (Hrsg.), Biographie und Psychologie (S. 108-116). Berlin: Springer.
- Thomae, H. (1988). Das Individuum und seine Welt (2., völlig neu bearbeitete Auflage). Göttingen: Hogrefe.
- Thurstone, L. L. (1938). Primary mental abilities. Psychometric Monographs, 1. Chicago: University of Chicago Press.
- Tinger, G. (1990). Das Experten- und Lehrsystem DSM-III-X: Ein Expertensystem für die klinische Diagnostik. Diagnostica, 36, 204-212.
- Tismer, K. G. (1976). Verhaltensbeobachtung bei Kindern und Jugendlichen. In K. Heller (Hrsg.), Handbuch der Bildungsberatung (Bd. 3, S. 817-836). Stuttgart: Klett.
- Todt, E. (1967). Differentieller Interessen-Test. Bern: Huber.
- Tomm, K. (1994). Die Fragen des Beobachters. Schritte zu einer Kybernetik zweiter Ordnung in der systemischen Therapie. Heidelberg: Carl Auer.
- Tränkle, U. (1983). Fragebogenkonstruktion. In H. Feger & J. Bredenkamp (Hrsg.), Enzyklopädie der Psychologie, Themenbereich B: Methodologie und Methoden, Serie I: Forschungsmethoden der Psychologie (Bd. 2: Datenerhebung, S. 222-301). Göttingen: Hogrefe.
- Tränkle, U. (1993). Fragebogen. In A. Schorr (Hrsg.), Handwörterbuch der Angewandten Psychologie (S. 243-248). Bonn: Deutscher Psychologen Verlag.
- Trebeck, R. (1961). Die Arbeitsplatzanalyse als Grundlage der Arbeitsplatzgestaltung, der Auswahl und Ausbildung von Mitarbeitern und Arbeitsbewertung. In K. Gottschaldt, Ph. Lersch, F. Sander & H. Thomae (Hrsg.), Handbuch der Psychologie in 12 Bänden (Bd.9, herausgegeben von A. Mayer & B. Herwig: Betriebspsychologie, S.210-243). Göttingen: Hogrefe.
- Trost, G., Blum, F., Fay, E., Hensgen, A., Klieme, E., Maichle, U. & Nauels, H. U. (1995). Test für medizinische Studiengänge (TMS): Studien zur Evaluation (19. Arbeitsbericht: 1. Februar 1994 bis 31. Januar 1995). Bonn-Bad Godesberg: Institut für Test- und Begabungsforschung.
- Truax, C. B. & Carkhuff, R. R. (1967). Toward effective counseling and psychotherapy training and practice. Chicago: Aldine Press.
- Ueckert, H. (1995). Expertensysteme. In W. Sarges (Hrsg.), Management-Diagnostik (2., vollständig überarbeitete und erweiterte Auflage, S. 789-796). Göttingen: Hogrefe.
- Ulich, D. (1982). Interaktionsbedingungen von Verbalisation. In G. L. Huber & H. Mandl (Hrsg.), Verbale Daten (S. 43-60). Weinheim: Beltz.
- Undeutsch, U. (1983). Exploration. In H. Feger & J. Bredenkamp (Hrsg.), Enzyklopädie der Psychologie. Themenbereich B: Methodologie und Methoden, Serie I: Forschungsmethoden der Psychologie (Bd. 2: Datenerhebung, S. 321-361). Göttingen: Hogrefe.
- Vennen, D. (1992). Behandlungsergebnisse und Wirkfaktoren von Eheberatung. Philosophische Dissertation, Bonn. Göttingen: Hogrefe.

- Völker, U. (1980). Grundlagen der Humanistischen Psychologie. In U. Völker (Hrsg.), *Humanistische Psychologie. Ansätze einer lebensnahen Wissenschaft vom Menschen* (S. 13-37). Weinheim: Beltz.
- Vrana, S., Mc Neil, D. W. & Mc Glynn, F. D. (1986). A structured interview for assessing dental fear. *Journal of Behavior, Therapy & Experimental Psychiatry*, 17, 175-178.
- Wagner, H. (1981). *Hamburger Verhaltensbeurteilungsliste (HAVEL)*. Göttingen: Hogrefe.
- Wagner, H. & Baumgärtel, G. (1978). *Hamburger Persönlichkeitsfragebogen für Kinder (HA-PEF-K)*. Handanweisung. Göttingen: Hogrefe.
- Wahl, D. (1982). Handlungsvalidierung. In G. L. Huber & H. Mandl (Hrsg.), *Verbale Daten* (S. 259-274). Weinheim: Beltz.
- Walsh, V. R. (1976). A test of the content and predictive validity of a structured interview. *Dissertation Abstracts*, 36 (12-A), 7965-7966.
- Walsh, W. B. (1967). Validity of self-report. *Journal of Counseling Psychology*, 14, 21-32.
- Wartegg, E. (o. J.). *Wartegg-Erzählungstests (WET)*. Göttingen: Hogrefe.
- Wartegg, E. (1939). Gestaltung und Charakter. *Zeitschrift für angewandte Psychologie. Beiheft* 84, 1-261.
- Wartegg, E. (1953). *Schichtdiagnostik. Der Wartegg-Zeichentest (WZT)*. (Einführung in die experimentelle Graphoskopie.) Göttingen: Hogrefe.
- Walter, H. J. (1996). *Angewandte Gestalttheorie in Psychotherapie und Psychohygiene*. Opladen: Westdeutscher Verlag.
- Warzecha, G. (1989). *Testverwaltungssystem EIDOS: Error-Identification-And-Description-of-Structure*, Version 6/1989. Frankfurt: Warzecha.
- Wechsler, D. (1964). *Hamburg-Wechsler-Intelligenztest für Erwachsene (HAWIE)* (3. Auflage). Deutsche Bearbeitung: A. Hardesty & H. Lauber. Herausgegeben von C. Bondy. Bern: Huber.
- Wechsler, D. (1991). *HAWIE-R*. Siehe Tewes (1991).
- Wehner, E.G. (Hrsg.). (1981). *Psychodiagnostik in Theorie und Praxis*. Frankfurt: Lang.
- Weidmann, M. (1987). *Zusammenhang zwischen zweimaliger Befragung von Medizinstudenten und Erfolg im Studium*. Unveröffentlichte Diplomarbeit, Bonn.
- Weinert, A. B. (1981). *Lehrbuch der Organisationspsychologie: Menschliches Verhalten in Organisationen*. München: Urban & Schwarzenberg.
- Weise, G. (1975). *Psychologische Leistungstests*. Göttingen: Hogrefe.
- Weiss, D.J. (1982). Improving measurement quality and efficiency with adaptive testing. *Applied Psychological Measurement*, 6, 473-492.
- Weiß, R. H. (1971). *Grundintelligenztest - Skala 3 (CFT 3)*. Deutsche Bearbeitung des Culture Fair Intelligence Test-Scale 3 von R. B. Cattell. Braunschweig: Westermann.
- Weiß, R. H. (1978). *Grundintelligenztest - Skala 2 (CFT 2)*. Deutsche Bearbeitung des Culture Fair Intelligence Test-Scale 2 von R. B. Cattell. Braunschweig: Westermann.
- Westhoff, K. & Kluck, M.-L. (1991). *Psychologische Gutachten schreiben und beurteilen*. Berlin: Springer.
- Westhofen, R. (1991). *Die methodologische Relevanz der Gütekriterien der klassischen Testtheorie für das diagnostische Gespräch*. Unveröffentlichte Psychologische Diplomarbeit. Bonn.
- Westmeyer, H. (1972). *Logik der Diagnostik. Grundlagen einer normativen Diagnostik*. Stuttgart: Kohlhammer.
- Westmeyer, H. (1993). *Persönlichkeitsdiagnostik*. In A. Schott (Hrsg.), *Handwörterbuch der Angewandten Psychologie* (S. 508-513). Bonn: Deutscher Psychologen Verlag.
- Wewetzer, K. H. (1979). *Psychologische Diagnostik*. Darmstadt: Wissenschaftliche Buchgesellschaft.

- Weyerer, S., Platz, S., Eichhorn, S., Mann, A., Ames, D. & Graham, N. (1988). Die deutsche Version des Brief Assessment Interviews (BAI): Ein Instrument zur Erfassung von Demenz und Depression. *Zeitschrift für Gerontopsychologie und -psychiatrie*, 1, 147-152.
- Weyerer, S., Geiger-Kabisch, C., Kröper, C., Denzinger, R. & Platz, S. (1990). Die Erfassung von Demenz und Depression mit Hilfe des Brief Assessment Interviews (BAI): Ergebnisse einer Reliabilitäts- und Validitätsstudie bei Altenheimbewohnern in Mannheim. *Zeitschrift für Gerontologie*, 23, 205-210.
- Wexley, K. N., Yuk, G. A., Kovacs, S. Z. & Sanders, R. E. (1972). Importance of contrast effects in employment interviews. *Journal of Applied Psychology*, 56, 45-48.
- Wexley, K. N., Sanders, R. E. & Yuk, G. A. (1973). Training interviewers to eliminate contrast effects in employment interviews. *Journal of Applied Psychology*, 57, 233-236.
- Wieczerkowski, W., Nickel, H., Janowski, A., Fittkau, B. & Rauer, W. (1974). Angstfragebogen für Schüler (AFS). Braunschweig: Westermann.
- Whyte, W. F. (1943). *Street corner society*. Chicago: University of Chicago Press.
- Wiggins, J. S. (1973). *Personality and prediction: Principles of personality assessment*. Reading, Mass.: Addison-Wesley.
- Wild, B. (1989). Neue Erkenntnisse zur Effizienz des „tailored“-adaptiven Testens. In K. D. Kubinger (Hrsg.), *Moderne Testtheorie* (S. 179-186). Weinheim: Beltz.
- Wildgrube, W. (1990). Computergestützte Diagnostik in einer Großorganisation. *Diagnostica*, 36, 127-147.
- Wilk, L. (1975). Die postalische Befragung. In K. Holm (Hrsg.), *Die Befragung 1* (S. 187-200). München: Francke.
- Wilks, S. S. (1946). Sample criteria for testing equality of means, equality of variances, and equality of covariances in normal multivariate distribution. *Annals of Mathematical Statistics*, 17, 257-281.
- Williams, J. B. W., Gibbson, M., First, M. B., Spitzer, R. L., Davies, M., Borus, J., Howes, M. J., Kane, J., Pope, H. G., Rounsaville, B. & Wittchen, H. U. (1992). The structured clinical interview for DSM-3-R (SCID) 2: Multisite test-retest reliability. *Archives of General Psychiatry*, 49, 630-636.
- Wilson, F. R., Genco, K. T. & Yager, G. G. (1985). Assessing the equivalence of paper-and-pencil vs. computerized tests: Demonstration of a promising methodology. *Computers in Human Behavior*, 1, 265-275.
- Winslow, G. S., Ballinger, B. R. & Mc Harg, A. M. (1985). Standardised psychiatric interview in elderly demented patients. *British Journal of Psychiatry*, 147, 545-546.
- Wittchen, H. U., Robins, L.N., Cottler, L. B., Sartorius, N., Burke, J. D., Regier, D. and Participants In The Multicentre WHO/ADAMHA Field Trails. (1991). Cross-cultural feasibility, reliability and source of variance of the composite international diagnostic interview (CID). *British Journal of Psychiatry*, 159, 645-653.
- Wittkowski, J. (1994). *Das Interview in der Psychologie. Interviewtechnik und Codierung von Interviewmaterial*. Opladen: Westdeutscher Verlag.
- Wittkowski, J. (1996). Zum aktuellen Status von Formdeutungsverfahren. *Diagnostica*, 42, 191-219.
- Woodworth, R. S. (1919). Examination of emotional fitness for warfare. *Psychological Bulletin*, 16, 59-60.
- Wottawa, H. (1980). *Grundriß der Testtheorie*. München: Juventa.
- Wottawa, H. & Hossiep, R. (1987). *Grundlagen psychologischer Diagnostik*. Göttingen: Hogrefe.
- Wright, B. D. (1977). Solving measurement problems with the Rasch model. *Journal of Educational Measurement*, 14, 97-116.
- Wright, B. D. & Masters, G. N. (1982). *Rating scale analysis*. Chicago: Mesa Press.

- Wright, B. D. & Stone, M. H. (1979). Best test design. Rasch measurement. Chicago: Mesa Press.
- Wuttke, J. (1980). Ziele und Probleme der Psychotherapieforschung. In W. Wittling (Hrsg.), Handbuch der Klinischen Psychologie (Bd.6: Klinische Psychologie in Forschung und Praxis, S. 16-41). Hamburg: Hoffman & Campe.
- Wyss, D. (1966). Die tiefenpsychologischen Schulen von den Anfängen bis in die Gegenwart. Entwicklung, Probleme, Krisen (2., durchgesehene und erweiterte und um ein Sachregister vermehrte Auflage). Göttingen: Vandenhoeck & Ruprecht.
- Yarrow, L. (1960). Interviewing children. In P. Mussen (Ed.), Handbook of research methods in Child Development (p. 561-602). New York: Harper & Row.
- Zeisel, H. (1965). Probleme der Aufschlüsselung. In R. König (Hrsg.), Das Interview (S. 290-318). Köln: Kiepenheuer & Witsch.
- Zetterberg, H. (1969). Methoden (der Soziologie: Das Interview). In R. König (Hrsg.), Soziologie (S. 194-224). (Umgearbeitete und erweiterte Neuausgabe). Fischer-Lexikon (FL Nr. 10). Frankfurt: Fischer.
- Zerssen, D. v. & Koeller, D. M. (1976). Beschwerdenliste. Weinheim: Beltz.
- Zielke, M. & Kopf-Mehnert, C. (1978). Veränderungsfragebogen des Erlebens und Verhaltens (VEV). Weinheim: Beltz.
- Ziler, H. (1975). Der Mann-Zeichen-Test in detail-statistischer Auswertung (2. Auflage). Münster: Aschendorff.
- Zulliger, H. (1952). Einführung in den Behn-Rorschach-Test (BERO-Test). (3. Auflage). Bern: Huber.
- Zulliger, H. (1955). Diapositiv-Z-Test. (Dia-Z-Test) (2. Auflage). Bern: Huber.
- Zuschlag, B. (1992). Das Gutachten des Sachverständigen. Rechtsgrundlagen, Fragestellungen, Gliederung, Rationalisierung. Göttingen: Verlag für Angewandte Psychologie.

Personenregister

- Abels, D. 269
ADAFI: Adaptiver Figurenfolgen-Lerntest 401
Aiken, L.R. 38
Algera, J.A. 247
Allehoff, W. 308
Althaus, D. 198, 238
Althoff, K. 99, 268-269, 393
Amelang, M. 4-6, 9, 19, 299, 338, 344, 351, 441
American Psychological Association (APA) 338
Amthauer, R. 39, 68-69, 99, 269, 271, 393, 421
Andersen, E. B. 174
Anger, H. 199, 214-215, 219, 221-222, 247, 252
Angleitner, A. 34, 36, 108, 111, 297, 299 300
Arbeitskreis Assessment Center 497
Arnold, W. 441
Asch, S.E. 200
Aschenbrenner, K. M. 376
Atkinson, J. W. 318
Atteslander, P. 247
Axhausen, S. 15

Bader, P. 406-407, 409
Bagozzi, R. P. 108, 111
Baker, F.B. 108, 111
Bales, R. F. 198
Bandura, A. 16, 106, 187, 190
Barendregt, J. T. 426
Bartel, H. 28
Bartenwerfer, H. 267
Bauer, W. 183, 189-190
Baumann, U. 6, 342
Baumgärte], F. 315
Baumgärtel, G. 44, 53
Bäumler, G. 68, 264, 269

Beaumont, J. G. 405-406
Becker, H. 222, 246, 427
Beckmann, D. 7, 53, 307, 311
Behn-Eschenburg, H. 323
Bellak, L. 325
Bellak, S. S. 325
Belt, J. A. 481
Benton, A. L. 463
Berg, M. 218
Berkowitz, L. 106, 187
Bemauer, F. 15
Berufsverband Deutscher Psychologen (BDP) 338, 431-432
Bierkens, P. 423
Binder, A. 315
Binet, A. 7, 422
Blase, H. 5, 212
Bach, D. 478-480
Bochenski, J. M. 186
Booth, J.F. 9, 394
Borkenau, P. 307
Bortz J. 28
Böttcher, H.R. 10, 13, 19, 151, 180, 219, 338
Brähler, E. 7, 53, 307
Brem-Gräser, L. 330, 423
Brickenkamp, R. 26, 119, 259, 267, 269, 307-308, 319, 330
Buggle, F. 315
Bühler, Ch. 14, 211, 270
Bukasa, B. 386, 397, 405
Bundesanstalt für Arbeit 397, 409
Bungard, W. 188, 199, 207
Burgoon, M. 415
Buss, D.M. 299
Byrne, B. M. 108, 111

Campbell, D. T. 108
Cannell, C.F. 220, 249

- Cantril, H. 219
 Cary 385
 Cattell, R. B. 7, 66, 118, 282, 284-285, 287, 299, 309-310
 Cierpka, M. 437
 Climent, C.E. 254
 Collins, M. 405
 Conrad, W. 31, 41, 66, 71, 76, 94-96, 101, 104-105, 108, 123, 151, 154, 169, 174, 177, 179-181, 267, 288
 Cook, St. W. 236
 Coyle, B.W. 108, 111
 Craig, K.H. 299
 Cranach, M. von 93, 187-188, 193, 198 199, 207-208
 Cronbach, L.J. 85, 252, 304, 341, 373-374
 Crossley, H.M. 247
 CSS (Complete Statistical System) 385
 Dahl, G. 37, 44
 Dahlstrom, L. 36
 Dahlstrom, W.G. 36, 304
 Dahmer, H. 219, 222, 245, 415
 Dahmer, J. 219, 222, 245, 415
 Dailey, C. A. 211
 Daniels, J. C. 40, 268
 Davidson, H. H. 323-324
 Deegener, G. 218
 Dehmelt, P. 198, 218, 238
 Deter, H.-C. W. 369
 Deutsch, M. 236
 Dienel, P.C. 380
 Dieterich, R. 55, 78, 84, 96, 99, 106, 113-114, 119, 121, 180-181, 188-189
 Dilling, H. 435
 Dirks, H. 484
 Dollard, J. 187
 Donati, R. 15
 Dony, M. 254
 Dörner, D. 390
 Dorsch, F. 4, 183, 265, 310, 346, 392, 493
 dpv (Deutscher Psychologen Verlag) 338, 426, 439-440, 447-418, 466
 Duhm, E. 198, 238
 Düker, H. 52, 269
 Durchholz, E. 10, 423, 483
 Eber, H.B. 309
 Eberwein, M. 218
 Edwards, A. L. 300, 304
 Edwards, W. 375
 Eggert, D. 112, 270, 298, 310
 Erven, H. 195
 Esser, M. 256, 313, 358
 Eysenck, H. J. 7, 14, 310
 Eysenck, M. W. 14
 Fahrenberg, J. 36, 66, 118-119, 304-305, 454, 464
 Farkas, M. G. 405
 Faßnacht, G. 71, 183, 188-189, 199, 207
 Fay, E. 7, 33, 122, 271, 291
 Feger, B. 133
 Fennekels, G. 189, 492, 494-495
 Fennekels, G.P. 97, 204, 491-492, 495-497, 499
 Fischer, G. 14, 54-55, 71, 79, 106, 123, 174
 Fiske, D. W. 108
 Fisseni, H. J. 33, 97, 189, 204, 211, 233, 236, 239, 247, 251, 254-256, 358-359, 471, 487, 491-492, 495-497, 499
 Föderation Deutscher Psychologenvereinigungen 288
 Fowler, F. J. 247
 Frances, A. 250
 Frank, L.K. 319
 Frenz, H. G. 198
 Frei, F. 479
 French, C. C. 405-406
 Frenz, H. G. 93, 187-188, 193, 199, 207-208
 Freud, S. 14
 Fricke, R. 15, 96-97, 131-132, 135, 137-138, 140-141, 145, 171, 391
 Frieling, E. 478-479, 481, 493-495, 498, 500
 Frinken, M. 247
 Fuchs, CH. 323
 Fuchs, W. 14, 359
 Funke, J. 390
 Funke, U. 358, 390, 478, 487, 500
 Fürntratt, E. 99
 Gatenby, E. V. 212
 Genco, K.T. 406
 Gielen, D. 247
 Giese, H. 214
 Gigerenzer, G. 151
 Glaser, R. 15
 Gleser, G.C. 85, 373-374

- Goffin, R. D. 108, 111
 Goldberg, L. 35
 Goldberg, L.R. 35-36
 Goldfried, M. R. 14, 360
 Göllner, D. 369
 Gösslbauer, J.P. 375, 381
 Graham, P. 250
 Graumann, C.F. 186, 303, 415, 474, 485
 Grawe, K. 15
 Green, B.F. 399
 Groffmann, K.J. 13, 19, 267, 319
 Gross, A. 218
 Gross, L.D. 252
 Grubitzsch, D. 14
 Guilford, J.P. 7, 42, 105, 215
 Guion, R.M. 96
 Gulliksen, H. 31
 Gunderson, E. K. E. 254
 Gunderson, J.G. 252
 Guthke, J. 10, 13, 19, 24, 151, 180, 219, 338, 401
 Guttman, L. 57

 Haas, R.M.P. 474
 Häcker, H. 66, 298
 Hageböck, J. 387, 395-401, 407
 Hagen, Cornelia von 236
 Haladyna, T. 38
 Halder-Sinn, P. 369-370
 Halsig, N. 33, 233, 236, 255, 358-359
 Halweg, K. 218
 Hampel, R. 36, 69, 454
 Hank, G. 218
 Hank, P. 405-406, 409
 Hartland, J. 189
 Hartmann, H.A. 9, 415, 417, 441, 471-472
 Hase, H.D. 35-36
 Hasemann, K. 184, 188, 197-199, 202, 207-208
 Hathaway, S. R. 36, 304, 398
 Haubl, R. 9, 188, 415, 441
 Heidenreich, K. 38-39
 Heil, F. E. 351
 Heiß, R. 441
 Helzer, J. E. 247
 Heneman, R. L. 487
 Hergovich, H. 409
 Hermans, H. 315
 Hess, U. 495-496, 498

 Hetzer, H. 270
 Hilke, R. 386, 397, 405, 408-409
 Hiltmann, H. 267
 Hirnrichs, J.R. 500
 Hodge, R.D. 252, 255
 Hoepfner, R. 7
 Hofer, P. J. 399
 Hofmann, K. 406-407, 409
 Honaker, L.M. 404, 406
 Hörn-rann, H. 15, 318
 Horn, W. 7, 36, 67-68, 99, 268, 271, 275-276, 282, 390, 393, 450
 Hornby, A. S. 5, 212
 Hornke, L.F. 385, 391-392, 394-395
 Hornthal, St. 483
 Horst, P. 79
 Hossiep, R. 6, 8-10, 19, 199, 338, 369, 374, 377-380
 Houwink, R.H. 325
 Hoyos, C. Graf 478-479
 Hron, A. 215, 219, 227
 Huber, G.L. 397
 Huber, H.P. 282, 287-288
 Huber, O. 183-184, 188, 199, 201
 Hubert, L.J. 108, 111
 Humboldt-Psychologie-Lexikon 5, 342
 Hurt, S. W. 250, 252, 255

 Irle, M. 308, 422
 Ittner, E. 33, 233, 236, 255, 358-359

 Jackson, D. N. 34, 36, 107
 Jackson, R. 131
 Jäger, A.O. 99, 268-269, 393
 Jäger, R. S. 4, 6, 8-9, 13, 19, 308, 338, 351, 354, 360-361, 363, 377, 384-385, 387-389, 394-395, 404, 407, 415, 420, 422, 441
 Jahoda, M. 236
 Janke, W. 342-343
 Jeserich, W. 491, 494-496
 Jessnitzer, K. 338
 Jochmann, W. 478, 482, 493, 495-496
 Joerger, K. 203
 Jüttemann, G. 14

 Kaden, S. 247
 Kaegi, A. 3, 212
 Kahn, R.L. 220, 249
 Kalinowsky-Czech, M. 254

- Kaminski, G. 5, 183, 189, 418
 Kamp, L.J. Th. van der 85
 Kapfer, E.L. 254
 Kasubek, W. 376
 Kazdin, A. E. 252
 Keil, W. 303
 Keller, G. 315
 Kelly, G. A. 14, 211
 Kemmler, L. 212
 Kent, R.N. 14, 360
 Keßler, B. H. 247, 441
 Kipnowski, A. 441
 Kirusek, T. J. 380
 Kisker, K. P. 415
 Kissner, R. 383, 386, 391-397
 Klann, N. 218
 Klapprott, J. 373, 381
 Klauer, K.J. 15, 33, 38, 85, 96, 129-131, 133-138, 144-147, 149, 270, 342, 344
 Klein, J. 9
 Kleinevoss, R. 477
 Klieme, E. 383, 385, 387, 389, 392-396, 399, 404-407, 409
 Klopfer, B. 323-324
 Kluck, M.-L. 9, 219-221, 227, 236, 247, 338, 441, 471
 Kluwe, R.H. 390, 401
 Kneubühler, H.U. 247
 Koch, K. 320, 330, 423, 459
 Kohli, M. 215
 Kolke, D.J. 252
 Kompa, A. 346, 348, 351, 354, 478, 482-483, 487
 Kopf-Mehnert, C. 308
 Kornadt, H. J. 320-321, 325, 327-328
 Kötter, S. 205
 Krampen, G. 107
 Kranz, H.T. 45, 84
 Kratzmeier, H. 268
 Krieger, W. 384, 388, 394-395
 Kristof, W. 282, 285-287
 Kriz, J. 15
 Kruse, A. 14
 Kubinger, K.D. 19, 151, 171, 386, 392-394, 405-409, 441
 Kuhnert, W. 198, 218, 238
 Kuliga, K. 385
 Küpper, Th. 247-248, 250, 252, 255
 Landy, F. J. 254
 Lane, J. W. 255
 Lang, A. 9, 441
 Langner, R. 401
 Langosch, I. 222, 246, 427
 Laplanche, J. 187, 318
 Lechler, P. 253
 Leeb, B. 250
 Lehr, U. 14, 212-213, 247, 255
 Lehrenkrauss, E. 483, 495
 Leichner, R. 13-15, 19, 129, 303, 317-318, 351, 354, 367, 413, 426
 Lennep, D.J. van 325
 Lennertz, E. 300
 Lewin, K. 186
 Liebel, H. 427
 Lienert, G.A. 26, 38, 41, 52, 54, 57, 60, 62, 65-66, 78, 84, 87, 106, 113, 119, 269, 326, 346, 348, 408
 Lindner, K. 140
 Lindzey, G. 327
 Links, P.S. 252
 Linstone, H.A. 378
 Lippert, S. 483, 487
 Lischer, S. 308
 Lohaus, A. 218
 Lord, F.M. 31, 71, 84, 91
 Lorenz, J. H. 204, 427
 Loretto, V. 232, 483
 Lück, H. E. 304-305
 Lutz, R. 219, 245, 247-248, 370
 Maccoby, E. E. 199, 215, 220-221, 247
 Maccoby, N. 199, 215, 220-221, 247
 Magnusson, D. 71
 Mai, N. 375
 Maier, O. 487
 Mailahn, N. 33, 233, 236, 255, 358-359
 Malle, B. F. 405, 407
 Marco, G. L. 404
 Martin, J.T. 393
 Mash, E.J. 218
 Mason, M.A. 481
 Masters, G. N. 174
 Matarazzo, J. D. 293, 399
 Maukisch, H. 478, 487, 500
 McBride, J. R. 393
 McClelland, M. C. 318
 McKinley, J.C. 36, 304, 398

- Meehl, P. E. 351, 354
 Mees, U. 97
 Meinefeld, W. 306
 Meyerhoff, H. 254
 Michel, L. 13, 19, 31, 41, 66, 71, 76, 94-96, 101, 104-105, 108, 123, 151, 154, 169, 174, 177, 179-181, 288, 319, 375
 Mische], W. 14, 16, 358
 Mittenecker, E. 300, 303-305
 Moser, K. 95, 149
 Müller, A. 464
 Müller, G.F. 376-377
 Mummendey, H. D. 299-305
 Münster, B. 308
 Murray, H.A. 14, 27, 107, 189, 211, 320, 325-326, 328, 455-456, 464
 Murstein, B.I. 319

 Nachreiner, F. 376-377
 Nanda, H. 85
 Neubauer, A. C. 405, 407
 Noelle, E. 219, 222, 236
 Nolen, P.A. 405
 Nordmann, E. 205
 Novick, M.R. 31, 71, 84, 91

 Olbrich, E. 358
 Odell, K. 405
 Olbrich, E. 33, 233, 236, 255-256, 359
 Ostendorf, F. 108, 111, 307

 Parry, H. J. 247
 Pawlik, K. 14, 96, 360
 Pelzmann, S. 426, 474
 Perez, M. 6
 Petermann, F. 4, 6-7, 9, 13, 16, 19, 198, 315, 338, 361
 Petermann, U. 7, 16, 315
 Pontalis, J.B. 187, 318
 Pryer, R.S. 319
 Pulver, U. 9, 441
 Quitmann, H. 15
 Raatz, U. 26, 38, 41, 54, 60, 62, 66, 78, 84, 87, 113, 119, 326, 346, 348, 408
 Rajaratnam, N. 85
 Rasch, G. 154, 174
 Rauch, M. 397
 Rauchfleisch, U. 218, 326
 Rausche, A. 301
 Reeb, W. 5, 212

 Rettig, K. 385, 392, 394-395
 Richter, H.E. 6-7, 53, 307, 427
 Ritz, B. 308
 Rock, D. L. 405
 Roest, F. 390
 Roidt, G.H. 38
 Ross, D. 190
 Ross, S.A. 190
 Ruch, W. 108, 111
 Ruffner, M. 415
 Rugg, D. 219
 Russell, J.T. 347
 Rutter, M. 250

 Saari, B. B. 108, 111
 Sarges, W. 219
 Sauermann, P. 266, 478
 Sawyer, J. 354
 Schaller, S. 14
 Scheiblechner, H. 174
 Scherer, K. R. 415
 Scheuch, E. K. 219, 222
 Scheurer, H. 360-361, 363
 Schmale, H. 282, 422
 Schmalt, H.-D. 422
 Schmid, F. W. 9, 441, 493
 Schmidt, G. 214
 Schmidt, K. J. 247
 Schmidt, L. R. 9, 247, 415, 441, 471, 473-474
 Schmidtchen, St. 9, 351, 373
 Schmidtke, A. 14
 Schmidtke, H. 422
 Schmitt, N. 108, 111
 Schmitt-Planert, A. 495-496, 498
 Schmitz-Scherzer, R. 255
 Schneewind, K. A. 188, 299, 309-310
 Schober, S. 317
 Scholz, O. B. 205, 218, 244, 314
 Schoppe, K. J. 68
 Schram], W. 199, 213, 247
 Schröder, G. 299, 309-310
 Schröder, R.D. 4
 Schuler, H. 358, 478, 495, 500
 Schulte, D. 245, 360
 Schwarzer, R. 315
 Schwenkmezger, P. 405-406, 409
 Seek, U. 247
 Sehringer, W. 330

- Seidenstücker, E. 247
 Seidenstücker, G. 247
 Seiffert, H. 414
 Seitz, W. 301, 351
 Selg, H. 36, 69, 183, 189-190, 454
 Sherman, R.E. 380
 Simon, T. 7, 422
 Sines, L.K. 247, 254
 Six, B. 306
 Sixtl, F. 27
 Skinner, B. F. 7
 Skre, I. 250
 Sloane, R.B. 370
 Solomon, J. 325
 Sommer, G. 52
 Sonnenberg, H.-G. 390, 401, 407
 Sonntag, K. 478, 481, 493-495, 498, 500
 Soskin, W. F. 254
 Spearman, C. 7, 267
 Spitznagel, A. 20, 188, 322-323, 417, 445
 Sprung, L. 10, 13, 19, 151, 180, 219, 338
 SPSS (Statistical Package for the Social Science) 385
 Staabs, G. von 204, 320, 331
 Staples, F.R. 370
 Starr, B. 325
 Stauffer, E. 315
 Steck, P. 326
 Stehle, W. 495
 Steinberg, M. 250, 256
 Stern, W. 6, 14
 Stevens, S. S. 28
 Stoll, F. 486-487
 Stone, M.H. 151-152, 154, 172-174
 Stults, D.M. 108, 111
 Stumpf, H. 34, 304, 383, 385, 387, 389, 392-396, 399, 404-407, 409
 Süllwold, F. 218
 Tanzer, N. 405
 Tatsuoaka, M.M. 309
 Taylor, H.C. 347
 Tent, L. 326
 Terdal, L. T. 218
 Testkuratorium der Föderation Deutscher Psychologenvereinigungen 407
 Tewes, U. 37, 119, 122, 264, 291, 299
 Thiel, R. 315
 Thomae, H. 6, 14, 186, 189, 211-212, 233, 255, 265, 441, 459, 461
 Thurstone, L. L. 7, 33, 36, 268, 271, 275
 Timaeus, E. 304-305
 Tinger, G. 399-400
 Tismer, K. G. 188, 199, 201
 Todt, E. 308, 422, 458
 Tomm, K. 219, 222, 245
 Tränkle, U. 31, 299-301, 310
 Trebeck, R. 478
 Trost, G. 7, 33, 359
 Trottmann-Geschwend, A. 315
 Truax, C.B. 370
 Turoff, M. 378
 Ueckert, H. 400
 Ulich, D. 213
 Undeutsch, U. 211, 236
 Urban, E. 405, 407
 Uslar, W. v. 427
 Vennen, D. 369, 503
 Völker, U. 15
 Vrana, S. 255
 Wagner, H. 44, 53, 198
 Wahl, D. 253, 308
 Waketield, H. 212
 Walbott, H. G. 415
 Walsh, V.R. 254
 Walsh, W.B. 254
 Walter, H. J. 437, 446, 469
 Wartegg, E. 330
 Warzecha, G. 385
 Weber, W. 397
 Wechsler, D. 67, 119, 122, 295, 463
 Wehner, E. G. 19
 Weidmann, M. 255
 Weinert, A. B. 480
 Weise, G. 5, 99, 124
 Weiss, D.J. 99, 391, 393
 Welsh, G. Sch. 36
 Wenninger, U. 386, 397
 Westhofen, R. 247-248, 250, 252, 255
 Westhoff, K. 9, 219-221, 227, 236, 247, 338, 441, 471
 Westmeyer, H. 4, 10, 341
 Weyerer, S. 250, 253, 255
 Whyte, W.F. 196
 Wieck, Th. 34

-
- | | |
|--|--|
| Wieczerkowski, W. 315 | Wurst, E. 171 |
| Wiggins, J. S. 297, 299-300, 351 | Wuttke, J. 24 |
| Wild, B. 393, 395 | Wyss, D. 15 |
| Wildgrube, W. 386-387, 395, 397, 405-406, 408-409 | Yager, G.G. 406 |
| Wilk, L. 214 | Yarrow, L. 218 |
| Wilks, S. S. 78 | |
| Williams, J. B. W. 250 | Zeidler, M. 483, 487 |
| Wilson, F.R. 406 | Zeisel, H. 236 |
| Windheuser, H. J. 370 | Zetterberg, H. 215 |
| Winslow, G.S. 250 | Zielinski, W. 4-6, 9, 19, 299, 315, 338, 344, 351, 441 |
| Wittchen, H.U. 250 | Zielke, M. 308 |
| Wittkowski, J. 227, 236, 321 | Ziler, H. 423 |
| Woodworth, R.S. 7 | Zinn, A. 198, 218, 238 |
| Wottawa, H. 6, 8-10, 19, 93, 111, 174, 176 178, 199, 300, 309, 338, 346, 369, 374, 377-380 | Zulliger, H. 323 |
| Wright, B. D. 151-152, 154, 172-174 | Zumkley, H. 320-321, 325, 327-328 |
| | Zuschlag, B. 338, 471 |

Sachregister

- Abweichungsnormen 114
 - ⇒ Variabilitätsnormen
- Act and React Testsystem (ART-90) 397
- Adaptive Tests 391
- Aktueller (u. biographischer) Ansatz 358
- Akzeptanz (einer Untersuchung) 423
- Alpha-Fehler (α) 344
- Analyse Qualitativer Daten (AQUAD) 397
- Anamnese 211-257
 - ⇒ Gespräch, Befragung, Exploration, Interview
- Angstfragebogen für Schüler (AFS) 315
- Anstrengungsvermeidungstest (AVT) 315
- Antwort, diagnostische 424-426
- Antworttendenzen (Fragebogen) 303-306
 - Ja-Sage-Tendenz 304
 - Kontrolle 210-213
 - Lügentendenz 211
 - Simulations- / Dissimulationstendenz 304
 - Soziale Erwünschtheit 304
- APA-Normen 338
 - (APA: American Psychological Association)
- Äquivalenznormen (Normierung) 113
- Assessment-Center 491
- Aufgabe (Testaufgabe) ⇒ Item 35, 38-40
- Augenscheinvalidität 244
- Aufmerksamkeits-Belastungs-Test (Test d2) 269
- Axiome der klassischen Testtheorie (KTT) 70
- Bales: Interaktionsschema 198
- Basisrate 346
- BDP-Normen 338
 - (BDP: Berufsverband Deutscher Psychologen)
- Baum-Test 330
- Bedeutsamkeit (Meßtheorie) 28, 364
- Befragung 312-257
 - ⇒ Gespräch, Anamnese, Exploration, Interview
- Begutachtung 439-475
 - ⇒ Gutachten
- Behn-Rorschach-Test (BERO) 323
- Beobachtung 183-209
 - ⇒ Verhaltensbeobachtung
- Beobachtungsbogen für Kinder im Vorschulalter (BBK) 198
- Berufs-Interessen-Test II (BIT II) 308
- Berufsordnung für Psychologen 8-9, 337-339
- Berufswahltest (BWT) 397
- Beta-Fehler (β) 344
- Bewerber-Selektion 477-489
- Binet (Binetarium) 7, 113
- Binomialmodell 144
- Biographischer (u. aktueller) Ansatz 358
- Biographisches Inventar zur Diagnose von Verhaltensstörungen (BIV) 308
- Bühler-Hetzer-Kleinkindertest (BHKT) 270
- Children's Apperception Test (CAT) 325
- Complete Statistical System (CSS) 385
- Computer Adaptives Testen (CAT) 397
- Computordiagnostik 383-410
 - Computersysteme 388
 - Computertests 389
 - Adaptive Tests 391
 - Einsatzfelder 401
 - Äquivalenz zwischen Papier-Bleistift-Test und ihren Computer-Versionen 404
- Computergesteuertes diagnostisches System auf normativer Grundlage (DIASYS) 398
- Cut-Off-Point 141
 - ⇒ Kritischer Punktwert
- Datenschutz 337-339
- Decisionsstudie (D-Studie) 85

- ⇒ Generalisierbarkeitstheorie
- Delphi-Methode 378
- Deskription (Fragebogen, Persönlichkeits-test) 297-299
- Deskription (u. Performanz): Untersuchungsebene 357
- Dezentrales Testvorgabe- und Auswertungssystem (DELTA) 386
- Diagnostic and Statistical Manual of Mental Disorders (DSM-III-X) 401
- Diagnostik
 - Aufbau 19-22
 - Aufgabenfelder 6-7, 13-16, 273, 503-506
 - Definition, Sachbedeutung 3-4, 19-21
 - Entstehungsgeschichte 6-16
 - Ethischer, juristischer Kontext 8-10, 337, 415, 441
 - Materiale Diagnostik 20, 503
 - Modellvorstellungen 13-16
- Diagnostische Situation 20, 415
- Diagnostische Untersuchung 20, 413-429
- Differentieller Interessen-Test (DIT) 308
- Differentieller Wissenstest (DWT) 99
- Diskriminante (u. konvergente) Trennschärfte 52
- Diskriminante (u. konvergente) Validität 109
- Dissimulations- / Simulationstendenz 304
- L' Echelle Metrique de l'Intelligente (Binet & Simon) 7
- Eichstichprobe ⇒ Normstichprobe 111, 120
- Eigenschaft (trait) u. Diagnostik 14, 273
- Eignung 265
- Elektronische Datenverarbeitung u. Diagnostik
 - ⇒ Computerdiagnostik 383-410
- Entscheidungstheorie 373-381
- Erfolgskontrolle 369-372
- Ethical Principles of Psychologists and Code of Conduct (APA) 338
- Ethischer, juristischer Kontext diagnostischen Vorgehens 8-10, 337, 408, 415, 441
- Exploration 211-257
 - ⇒ Gespräch, Anamnese, Befragung, Interview
- Externale Konstruktionsstrategie 35
 - Test, Fragebogen
- Eysenck Personality Inventory (EPI) 112
- Fähigkeit 265
- Fehler der Informationsverarbeitung 199-201
 - ⇒ Verzerrungstendenzen
- Fertigkeit 265
- Formdeutungsverfahren 317-333
- Freiburger Persönlichkeitsinventar (FPI, FPI-R) 36, 305, 306
- Fragebogen 297-316
 - ⇒ Persönlichkeitstest
- Fragebogen zur Erfassung von Aggressivitätsfaktoren (FAF) 69
- Fragebogen zur Kontrollüberzeugung (IPC) 107
- Fremdbestimmung
 - ⇒ Selbstbestimmung 9-10
- Fuchs-Rarschach-Test (FURO) 323
- Geist-Bilder-Interessen-Inventar (GBII) 315
- Generalisierbarkeits-Studie (G-Studie) 85
 - ⇒ Decisionsstudie
- Generalisierbarkeitstheorie 85
- Generative Regeln (Kriteriumsorientierte Tests) 133
- Gespräch ⇒ Anamnese, Befragung, Exploration, Interview
 - Abgrenzung (Definition) 212
 - Arten / Klassifikation 214
 - Auswertung 236
 - Wiedergabe des Originalgesprächs 237
 - Zusammenfassung des Gesprächs 237
 - Schematische Z. 237
 - Thematische Z. 237, 239
 - Beitrag zu Diagnostik und Intervention 243
 - Durchführung 233
 - Fehler 247
 - Fragetechniken 219
 - Gütekriterien 247
 - Vorbereitung 226
- Gestalterische Verfahren 317, 330
 - ⇒ Projektive Verfahren
- Gießen-Test (GT) 51, 53, 307, 311
- Grundintelligenztest - Skala 3 (CFT 3) 99
- Grundkenntnisse, diagnostische 23
- Gruppentest für die soziale Einstellung (SET) 203
- Gutachten, psychologisches 439-475
 - Befund 456
 - Befundliste 462, 464
 - Befundskizze 463, 465

- Gliederung 441
- Richtlinien (BDP) 338, 439, 467
- Stellungnahme 466
- Tätigkeitsfelder 439
- Untersuchungsbericht 447
- Vorgeschichte 444
- Gütekriterien (Test) 66-110
- Objektivität 67
- Reliabilität 70
- Validität 93
- Guthnan-Skala 57
- Hamburg-Wechsler-Intelligenztest für Erwachsene (HAWIE, HAWIE-R) 37, 67, 122, 299
- Hamburger Neurotizismus- und Extraversionskala (HANES, KJ) 315
- Hamburger Persönlichkeitsfragebogen für Kinder (HAPEF-K) 53
- Hamburger Verhaltensbeurteilungsliste (HAWEL) 198
- Heutismen (aus projektiven Verfahren) 324, 329, 332
- Hilfslosigkeits- und Selbstwirksamkeitsskala 315
- Hochrechnung 81-84
 - ⇒ Spearman-Brown-Formula
- Homogenität 54-59
- Faktorenanalyse 56
- Guttman-Skala 57
- Interkorrelation 55
- Rasch-Modell 59, 174
- Hypothesen 414
- Index der kategorialen Häufigkeit 41
 - ⇒ Schwierigkeit
- Integrative Diagnostik 411, 413-501
 - ⇒ Synthese in der Diagnostik
- Intelligenz(funktionen) 267
- Intelligenzmodelle 267
- Intelligenz-Struktur-Test (IST 70) 39, 69, 90, 93, 99, 393
- Interaktionistische Persönlichkeitspsychologie u. Diagnostik 16
- Internale Konstruktionsstrategie 36
 - ⇒ Test
- Interpretation 414
- Intervallskala 28, 123
- Intervention
 - Definition 5
 - Modelle 13-16
 - Beiträge zur Intervention
 - Klassische Testtheorie 122
 - Kriteriumsorientierte Testtheorie 148
 - Rasch-Modell 179
 - Verhaltensbeobachtung 203
 - Gespräch (Exploration, Interview, Anamnese) 243
 - Leistungstests 270
 - Persönlichkeitstest 313
 - Projektive Verfahren 321
 - Computerdiagnostik 401
 - Diagnostischer Prozeß 426
 - Therapieplanung 428
- Interview 211-257
 - ⇒ Gespräch, Anamnese, Befragung, Interview
- Intuitiv / rationale Konstruktionsstrategie 211-257
 - ⇒ Test, Fragebogen
- Item (klassische Testtheorie) 35, 38-40, 40-65
- Analyse (Itemanalyse) 40-65
 - Homogenität 54
 - Schwierigkeit 41
 - Trennschärfe 47
- Arten 38-40
- Generierung 38
- Selektion / Testrevision 59-65
- Item-Charakteristik-Kurve (ICC) 153, 175
- Itemparameter (Rasch) 151
 - ⇒ Personenparameter (Rasch)
- Juristischer, ethischer Kontext diagnostischen Vorgehens 8-10, 337, 415, 441
- Kinder: Gespräche mit K. 218
- Klassifikation (u. Selektion) 341-349
- Klassifikationsfehler 344
 - ⇒ Alpha-, Betafehler
- Klassische Testtheorie (KTT) 31-127
- Klinisch-psychologische / psychotherapeutische Intervention: Leitsätze zur Dokumentation (BDP) 338
- Klinische Urteilsbildung 351-355
 - ⇒ Statistische Urteilsbildung
- Kontrastfehler 201
- Konstruktionsstrategien (Test, Fragebogen)
 - (intuitiv / rational, external, internal)

- Konvergente (u. diskriminante) Trennschärfe 52
- Konvergente (u. diskriminante) Validität 109
- Konzentrationstests 269
- Konzentrations-Leistungs-Test (KLT) 269
- Konzentrations-Verlaufs-Test (KVT) 269
- Korrespondenzprobleme (Untersuchung) 421
- Kriteriumsorientierte Leistungsmessung 129
- Kritische Differenz (zwischen Testscores) 92
- Kritik der klassischen Testtheorie 123
- Kritischer Punktwert (Cut-Off-Point) 141
- „Leerer Stuhl“ (Gestalttherapeutisches Verfahren) 438
- Leipziger Testsystem 396
- Leistung(sfunktionen) 266
- Leistungsmotivationstest (LMT) 315
- Leistungs-Prüf-System (LPS) 275, 393, 450
- Leistungstest 263-296
 - Allgemeine 267
 - Spezielle 269
- Lern- und Gedächtnistest (LGT 3) 68, 269
- Lincoln-Oseretzky-Scale, Kurzform (LOS KF 18) 270
- Logistische Funktion (Rasch-Skala) 154
- Lokale stochastische Unabhängigkeit 176
- Lügenskala 304
 - ⇒ Offenheit / Verslossenheit
- MAILBOX-90 390
- Messen 27-28
- Mildefehler 200
- Minderungskorrektur (Test, Validität) 102-105
 - Einfache 102
 - Doppelte 103
- Minnesota Multiphasic Personality Inventory (MMPI) 36, 306
- Modellvorstellungen (Diagnostik, Intervention) 13-17
- Multi-Attributive-Utility-Theory (MAUT) 172
- Multimethodale / Multimodale Diagnostik 411, 413, 431, 439, 477, 491
- Multiple choice (Mehrfachauswahl) 38
- Multitrait-multimethod-Validierung 108
- Nebengütekriterien 66
 - (Normierung / Ökonomie / Nützlichkeit / Vergleichbarkeit)
- NEO-Fünf-Faktoren Inventar (NEO-FFI) 307
- Nominalskala 28-29
- Normative Diagnostik 9
- Normen (Test) 111, 120
 - Äquivalenznormen 113
 - Eichstichprobe 120
 - kulturell-ethische Normen 122
 - Normalverteilung 120
 - Normieren 111
 - Probleme 120
 - Prozentränge 115
 - Rohwerte 111
 - Stichprobenabhängigkeit 121
 - Transformierte Werte 112
 - Übliche Normskalen 119
 - Variabilitätsnormen 114
- Normstichprobe 120
- Nutzen diagnostischen Vorgehens 373-382
 - ⇒ Utilität
- Object Relation Technique (ORT) 325
- Objektivität (Test) 66-69
 - Auswerterobjektivität 68
 - Durchführungsobjektivität 67
 - Interpretationsobjektivität 68
 - Probleme 69
 - Spezifische (Rasch-Skala) 177
- Offenheit / Verslossenheit 304
 - ⇒ (Fragebogen: Antworttendenz)
- Ordinalskala 28, 123, 179
- Papier-Bleistift-Tests und ihre Computer-Versionen 389
- Performanz (u. Deskription) 297-299
 - ⇒ Fragebogen, Persönlichkeitstest
- Personality Research Form (PRF) 33, 304
- Personenparameter (Rasch) 151-152
 - ⇒ Itemparameter (Rasch)
- Persönlichkeits-Entfaltungs-Verfahren 317-333
 - ⇒ Projektive Verfahren
- Persönlichkeitsinventar 317-333
 - ⇒ Persönlichkeitstest, Fragebogen
- Persönlichkeitspsychologie u. Diagnostik 13-16
- Persönlichkeitstest 317-333
 - ⇒ Fragebogen, Persönlichkeitsinventar
- Picture Frustration Test (PFT) 119, 326
- Plazierung 343

- Positionseffekt 200
- Probabilistische Testtheorie (Rasch) 151-182
- Profilanalyse (Testscores) 275
- Profilvergleich (Testscores) 282
- Projektion / projektiv 317-319
 - Identifikation 327
 - Klassifikation 319-320
- Projektive Verfahren 317-333
 - Umschreibung (Definition) 317
 - Arten / Klassifikation
 - Formdeutungsverfahren 322
 - Verbal-thematische Verfahren 325
 - Zeichnerische u. gestalterische Verfahren 330
- Beitrag zu Diagnostik und Intervention 321
- Probleme 320
- Prozentränge (Normierung) 115
- Prozeßdiagnostik 360
 - ⇒ Statusdiagnostik
- Prozeßorientierung u. Diagnostik 14
- Prüfsystem für Schul- und Bildungsberatung (PSB) 99
- Psychodynamischer Ansatz u. Diagnostik 15
- Rasch-Modell (probabilistische Testtheorie) 151-182
- Rationalskala 28-29
 - ⇒ Verhältnisskala
- Reagibilität 24
- Rechnergestütztes Psychodiagnostisches System (RPS) 396
- Reduzierter Wechsler-Intelligenztest (WIP) 37
- Reliabilität (Test) 72-92
 - Axiome der klassischen Testtheorie 70
 - Definition 72
 - Generalisierbarkeitstheorie 85
 - Halbierungsreliabilität 79
 - Hochrechnung 81
 - Inkompatibilität mit Validität 104
 - Konsistenz 86
 - Paralleltestreliabilität 78
 - Retestreliabilität 76
 - Spearman-Brown-Formula 81
- Reliabilitätsindex 101
- Rorschach 322
 - ⇒ Formdeutverfahren
- Rückmeldung 484
 - ⇒ Bewerberselektion
- School Apperception Method (SAM) 325
- Steno-Test 331
- Schwierigkeit(sindex: Test, Item) 41-46
 - Mehrstufige Antworten 43
 - Zweistufige Antworten 41
- Schweigepflicht 8-10, 337, 415, 441
 - ⇒ Vertraulichkeit 8-10, 337, 415, 441
- Sechzehn-Persönlichkeitsfaktoren-Test (16 PF) 299
- Selbstbeschreibung 297-299
 - (Fragebogen, Persönlichkeitstest)
- Selbstbeschreibung u. Verhalten 301-306
- Selbstbestimmung 9-10
 - ⇒ Fremdbestimmung
- Selektion (u. Klassifikation) 351-355
- Selektionrate 346
- Self-fulfilling prophecy 201
 - ⇒ Erwartung als Fehler
- Senior Apperception Technique (SAT) 325
- Simulations- / Dissimulationstendenz 304
- Sixteen Personality Factor Questionnaire (16 PF) 118
- Skalenniveau 28
- Spearman-Brown-Formula (Halbierungsreliabilität) 80, 83
- Social Desirability-Scale Edwards (SDS-E) 305
- Soziale Erwünschtheit (Antworttendenz) 304
- Sozialer Kontext der Diagnostik 8-9, 337-339, 415-418
- Split-half reliability 79
 - ⇒ Halbierungsreliabilität
- Standardisiert / Standardisierung 66
- Standardmeßfehler (Test) 90
- Standardschätzfehler (Test) 91
- Statistical Package for the Social Sciences (SPSS) 385
- Statistische (u. klinische) Urteilsbildung 391-355
- Statusdiagnostik 360
 - ⇒ Prozeßdiagnostik
- Synthese in der Diagnostik 411, 413-501
 - ⇒ Integrative Diagnostik
- Szenario-Technik 379
- Taylor-Russel-Tafeln 347
- Test
 - Arten 26, 37
 - Definition 24
 - Gütekriterien 66-108

- Itemanalyse 40-65
- Klassifikation 26, 37, 267-270
- Konstruktion
 - Klassische Testtheorie 31-127
 - Kriteriumsorientierte Leistungsmessung 129-150
 - Raschmodell 151-182
- Konstruktionsstrategien 35
- Kritik der klassischen Testtheorie 123
- Kritische Differenz 92
- Normierung 111
- Testrevision 59
- Vertrauensbereich 89
- Testbewertung
 - Kriterien 288
 - Beispiel (HAWIE-R) 291
- Test-Merkmal: Abgrenzung (Definition) 32
- Testtheorien
 - Klassische 31-127
 - Kriteriumsorientierte 129-150
 - Probabilistische (Rasch) 151-182
- Thematischer Apperzeptionstest (TAT) 325
- Thematischer Gestaltungstest (TGT-S) 325
- Trait u. Diagnostik 14, 273
 - ⇒ Eigenschaft
- Trennschärfe (Test, Item) 47-54
- Übersetzungsprobleme 418
 - (Proband / Untersucher)
- Ü-Koeffizient 135, 138
- Untersuchungsverlauf 411-429
- Urteilsbildung
 - Computer-gesteuert 401
 - Integration 424, 456, 466
 - Klassifikation 342
 - Selektion 245
 - Statistische, klinische 351
- Utilität diagnostischen Vorgehens 373-382
 - ⇒ Nutzen
- Validität 94-111
 - Abgrenzung (Definition) 94
 - Arten 95
 - Grenzen 101
 - Handlungs- 253, 255
 - Inkompatibilität (partielle) mit Reliabilität 104
 - Inhaltsvalidität / Kontentvalidität 95
 - Kommunikative 254
 - Konstruktvalidität 105
 - Konvergente / diskriminante 110
 - Kriteriumsbezogene 98
 - Minderungskorrektur (einfache / doppelte) 102-103
 - Multitrait-Multimethod-Validierung 108
 - Reliabilitätsindex 101
 - Übereinstimmungsvalidität 98
 - Vorhersagevalidität 100
- Variabilitätsnormen 114
 - ⇒ Normen / Normierung
- Variablen (Definition) 23
- Verbal-thematische Verfahren 325
- Varianten 325
- Verhaltensbeobachtung ⇒ Beobachtung
- Abgrenzung (Definition) 183
- Anwendung in der Diagnostik 203
 - (begleitend / selbständig)
- Beitrag zu Diagnostik und Intervention 203
- Einheiten bilden 187
- Einteilung 193
- Fehler
 - Allgemein 199
 - Speziell 201
- Fixierung (mechanisch / symbolisch) 197
- Gütekriterien 206
- Vorteile / Nachteile 206
- Veränderungsmessung 361
- Aporien 361
 - Bedeutsamkeitsproblem 364
 - Regression zur Mitte 365
 - Reliabilitäts-Validitäts-Dilemma
- Verhaltenstherapeutische Diagnostik: Kompendium 218
- Verhältnisskala 28-29
 - ⇒ Rationalskala
- Vertrauensbereich (eines Testscores) 89
- Vertraulichkeit 8-10, 337, 415, 441
- Verzerrungstendenzen 199-202
 - Erwartung (self-fulfilling prophecy) 201
 - Hofeffekt (halo effekt) 200
 - Kontrastfehler 202
 - Milde 200
 - Position 200
 - Strenge 200
 - Überstrahlungseffekt 200
 - Zentrale Tendenz 200
- Vier-Bilder-Test 492

Wartegg-Zeichen-Test (WZT) 330
Wiener-Testsystem 396
Wilde-Intelligenz-Test (WIT) 268

Zeichnerische Verfahren 317, 330
 \Rightarrow Projektive Verfahren
Zufallskorrektur (Schwierigkeitsindex) 42

Personenregister

- Abels, D. 269
ADAFI: Adaptiver Figurenfolgen-Lerntest 401
Aiken, L.R. 38
Algera, J.A. 247
Allehoff, W. 308
Althaus, D. 198, 238
Althoff, K. 99, 268-269, 393
Amelang, M. 4-6, 9, 19, 299, 338, 344, 351, 441
American Psychological Association (APA) 338
Amthauer, R. 39, 68-69, 99, 269, 271, 393, 421
Andersen, E. B. 174
Anger, H. 199, 214-215, 219, 221-222, 247, 252
Angleitner, A. 34, 36, 108, 111, 297, 299 300
Arbeitskreis Assessment Center 497
Arnold, W. 441
Asch, S.E. 200
Aschenbrenner, K. M. 376
Atkinson, J. W. 318
Atteslander, P. 247
Axhausen, S. 15

Bader, P. 406-407, 409
Bagozzi, R. P. 108, 111
Baker, F.B. 108, 111
Bales, R. F. 198
Bandura, A. 16, 106, 187, 190
Barendregt, J. T. 426
Bartel, H. 28
Bartenwerfer, H. 267
Bauer, W. 183, 189-190
Baumann, U. 6, 342
Baumgärte], F. 315
Baumgärtel, G. 44, 53
Bäumler, G. 68, 264, 269

Beaumont, J. G. 405-406
Becker, H. 222, 246, 427
Beckmann, D. 7, 53, 307, 311
Behn-Eschenburg, H. 323
Bellak, L. 325
Bellak, S. S. 325
Belt, J. A. 481
Benton, A. L. 463
Berg, M. 218
Berkowitz, L. 106, 187
Bemauer, F. 15
Berufsverband Deutscher Psychologen (BDP) 338, 431-432
Bierkens, P. 423
Binder, A. 315
Binet, A. 7, 422
Blase, H. 5, 212
Bach, D. 478-480
Bochenski, J. M. 186
Booth, J.F. 9, 394
Borkenau, P. 307
Bortz J. 28
Böttcher, H.R. 10, 13, 19, 151, 180, 219, 338
Brähler, E. 7, 53, 307
Brem-Gräser, L. 330, 423
Brickenkamp, R. 26, 119, 259, 267, 269, 307-308, 319, 330
Buggle, F. 315
Bühler, Ch. 14, 211, 270
Bukasa, B. 386, 397, 405
Bundesanstalt für Arbeit 397, 409
Bungard, W. 188, 199, 207
Burgoon, M. 415
Buss, D.M. 299
Byrne, B. M. 108, 111

Campbell, D. T. 108
Cannell, C.F. 220, 249

- Cantril, H. 219
 Cary 385
 Cattell, R. B. 7, 66, 118, 282, 284-285, 287, 299, 309-310
 Cierpka, M. 437
 Climent, C.E. 254
 Collins, M. 405
 Conrad, W. 31, 41, 66, 71, 76, 94-96, 101, 104-105, 108, 123, 151, 154, 169, 174, 177, 179-181, 267, 288
 Cook, St. W. 236
 Coyle, B.W. 108, 111
 Craig, K.H. 299
 Cranach, M. von 93, 187-188, 193, 198 199, 207-208
 Cronbach, L.J. 85, 252, 304, 341, 373-374
 Crossley, H.M. 247
 CSS (Complete Statistical System) 385
 Dahl, G. 37, 44
 Dahlstrom, L. 36
 Dahlstrom, W.G. 36, 304
 Dahmer, H. 219, 222, 245, 415
 Dahmer, J. 219, 222, 245, 415
 Dailey, C. A. 211
 Daniels, J. C. 40, 268
 Davidson, H. H. 323-324
 Deegener, G. 218
 Dehmelt, P. 198, 218, 238
 Deter, H.-C. W. 369
 Deutsch, M. 236
 Dienel, P.C. 380
 Dieterich, R. 55, 78, 84, 96, 99, 106, 113-114, 119, 121, 180-181, 188-189
 Dilling, H. 435
 Dirks, H. 484
 Dollard, J. 187
 Donati, R. 15
 Dony, M. 254
 Dörner, D. 390
 Dorsch, F. 4, 183, 265, 310, 346, 392, 493
 dpv (Deutscher Psychologen Verlag) 338, 426, 439-440, 447-418, 466
 Duhm, E. 198, 238
 Düker, H. 52, 269
 Durchholz, E. 10, 423, 483
 Eber, H.B. 309
 Eberwein, M. 218
 Edwards, A. L. 300, 304
 Edwards, W. 375
 Eggert, D. 112, 270, 298, 310
 Erven, H. 195
 Esser, M. 256, 313, 358
 Eysenck, H. J. 7, 14, 310
 Eysenck, M. W. 14
 Fahrenberg, J. 36, 66, 118-119, 304-305, 454, 464
 Farkas, M. G. 405
 Faßnacht, G. 71, 183, 188-189, 199, 207
 Fay, E. 7, 33, 122, 271, 291
 Feger, B. 133
 Fennekels, G. 189, 492, 494-495
 Fennekels, G.P. 97, 204, 491-492, 495-497, 499
 Fischer, G. 14, 54-55, 71, 79, 106, 123, 174
 Fiske, D. W. 108
 Fisseni, H. J. 33, 97, 189, 204, 211, 233, 236, 239, 247, 251, 254-256, 358-359, 471, 487, 491-492, 495-497, 499
 Föderation Deutscher Psychologenvereinigungen 288
 Fowler, F. J. 247
 Frances, A. 250
 Frank, L.K. 319
 Frenz, H. G. 198
 Frei, F. 479
 French, C. C. 405-406
 Frenz, H. G. 93, 187-188, 193, 199, 207-208
 Freud, S. 14
 Fricke, R. 15, 96-97, 131-132, 135, 137-138, 140-141, 145, 171, 391
 Frieling, E. 478-479, 481, 493-495, 498, 500
 Frinken, M. 247
 Fuchs, CH. 323
 Fuchs, W. 14, 359
 Funke, J. 390
 Funke, U. 358, 390, 478, 487, 500
 Fürntratt, E. 99
 Gatenby, E. V. 212
 Genco, K.T. 406
 Gielen, D. 247
 Giese, H. 214
 Gigerenzer, G. 151
 Glaser, R. 15
 Gleser, G.C. 85, 373-374

- Goffin, R. D. 108, 111
 Goldberg, L. 35
 Goldberg, L.R. 35-36
 Goldfried, M. R. 14, 360
 Göllner, D. 369
 Gösslbauer, J.P. 375, 381
 Graham, P. 250
 Graumann, C.F. 186, 303, 415, 474, 485
 Grawe, K. 15
 Green, B.F. 399
 Groffmann, K.J. 13, 19, 267, 319
 Gross, A. 218
 Gross, L.D. 252
 Grubitzsch, D. 14
 Guilford, J.P. 7, 42, 105, 215
 Guion, R.M. 96
 Gulliksen, H. 31
 Gunderson, E. K. E. 254
 Gunderson, J.G. 252
 Guthke, J. 10, 13, 19, 24, 151, 180, 219, 338, 401
 Guttman, L. 57

 Haas, R.M.P. 474
 Häcker, H. 66, 298
 Hageböck, J. 387, 395-401, 407
 Hagen, Cornelia von 236
 Haladyna, T. 38
 Halder-Sinn, P. 369-370
 Halsig, N. 33, 233, 236, 255, 358-359
 Halweg, K. 218
 Hampel, R. 36, 69, 454
 Hank, G. 218
 Hank, P. 405-406, 409
 Hartland, J. 189
 Hartmann, H.A. 9, 415, 417, 441, 471-472
 Hase, H.D. 35-36
 Hasemann, K. 184, 188, 197-199, 202, 207-208
 Hathaway, S. R. 36, 304, 398
 Haubl, R. 9, 188, 415, 441
 Heidenreich, K. 38-39
 Heil, F. E. 351
 Heiß, R. 441
 Helzer, J. E. 247
 Heneman, R. L. 487
 Hergovich, H. 409
 Hermans, H. 315
 Hess, U. 495-496, 498

 Hetzer, H. 270
 Hilke, R. 386, 397, 405, 408-409
 Hiltmann, H. 267
 Hirnrichs, J.R. 500
 Hodge, R.D. 252, 255
 Hoepfner, R. 7
 Hofer, P. J. 399
 Hofmann, K. 406-407, 409
 Honaker, L.M. 404, 406
 Hörn-rann, H. 15, 318
 Horn, W. 7, 36, 67-68, 99, 268, 271, 275-276, 282, 390, 393, 450
 Hornby, A. S. 5, 212
 Hornke, L.F. 385, 391-392, 394-395
 Hornthal, St. 483
 Horst, P. 79
 Hossiep, R. 6, 8-10, 19, 199, 338, 369, 374, 377-380
 Houwink, R.H. 325
 Hoyos, C. Graf 478-479
 Hron, A. 215, 219, 227
 Huber, G.L. 397
 Huber, H.P. 282, 287-288
 Huber, O. 183-184, 188, 199, 201
 Hubert, L.J. 108, 111
 Humboldt-Psychologie-Lexikon 5, 342
 Hurt, S. W. 250, 252, 255

 Irle, M. 308, 422
 Ittner, E. 33, 233, 236, 255, 358-359

 Jackson, D. N. 34, 36, 107
 Jackson, R. 131
 Jäger, A.O. 99, 268-269, 393
 Jäger, R. S. 4, 6, 8-9, 13, 19, 308, 338, 351, 354, 360-361, 363, 377, 384-385, 387-389, 394-395, 404, 407, 415, 420, 422, 441
 Jahoda, M. 236
 Janke, W. 342-343
 Jeserich, W. 491, 494-496
 Jessnitzer, K. 338
 Jochmann, W. 478, 482, 493, 495-496
 Joerger, K. 203
 Jüttemann, G. 14

 Kaden, S. 247
 Kaegi, A. 3, 212
 Kahn, R.L. 220, 249
 Kalinowsky-Czech, M. 254

- Kaminski, G. 5, 183, 189, 418
 Kamp, L.J. Th. van der 85
 Kapfer, E.L. 254
 Kasubek, W. 376
 Kazdin, A. E. 252
 Keil, W. 303
 Keller, G. 315
 Kelly, G. A. 14, 211
 Kemmler, L. 212
 Kent, R.N. 14, 360
 Keßler, B. H. 247, 441
 Kipnowski, A. 441
 Kirusek, T. J. 380
 Kisker, K. P. 415
 Kisser, R. 383, 386, 391-397
 Klann, N. 218
 Klapprott, J. 373, 381
 Klauer, K.J. 15, 33, 38, 85, 96, 129-131, 133-138, 144-147, 149, 270, 342, 344
 Klein, J. 9
 Kleinevoss, R. 477
 Klieme, E. 383, 385, 387, 389, 392-396, 399, 404-407, 409
 Klopfer, B. 323-324
 Kluck, M.-L. 9, 219-221, 227, 236, 247, 338, 441, 471
 Kluwe, R.H. 390, 401
 Kneubühler, H.U. 247
 Koch, K. 320, 330, 423, 459
 Kohli, M. 215
 Kolke, D.J. 252
 Kompa, A. 346, 348, 351, 354, 478, 482-483, 487
 Kopf-Mehnert, C. 308
 Kornadt, H. J. 320-321, 325, 327-328
 Kötter, S. 205
 Krampen, G. 107
 Kranz, H.T. 45, 84
 Kratzmeier, H. 268
 Krieger, W. 384, 388, 394-395
 Kristof, W. 282, 285-287
 Kriz, J. 15
 Kruse, A. 14
 Kubinger, K.D. 19, 151, 171, 386, 392-394, 405-409, 441
 Kuhnert, W. 198, 218, 238
 Kuliga, K. 385
 Küpper, Th. 247-248, 250, 252, 255
 Landy, F. J. 254
 Lane, J. W. 255
 Lang, A. 9, 441
 Langner, R. 401
 Langosch, I. 222, 246, 427
 Laplanche, J. 187, 318
 Lechler, P. 253
 Leeb, B. 250
 Lehr, U. 14, 212-213, 247, 255
 Lehrenkrauss, E. 483, 495
 Leichner, R. 13-15, 19, 129, 303, 317-318, 351, 354, 367, 413, 426
 Lennep, D.J. van 325
 Lennertz, E. 300
 Lewin, K. 186
 Liebel, H. 427
 Lienert, G.A. 26, 38, 41, 52, 54, 57, 60, 62, 65-66, 78, 84, 87, 106, 113, 119, 269, 326, 346, 348, 408
 Lindner, K. 140
 Lindzey, G. 327
 Links, P.S. 252
 Linstone, H.A. 378
 Lippert, S. 483, 487
 Lischer, S. 308
 Lohaus, A. 218
 Lord, F.M. 31, 71, 84, 91
 Lorenz, J. H. 204, 427
 Loretto, V. 232, 483
 Lück, H. E. 304-305
 Lutz, R. 219, 245, 247-248, 370
 Maccoby, E. E. 199, 215, 220-221, 247
 Maccoby, N. 199, 215, 220-221, 247
 Magnusson, D. 71
 Mai, N. 375
 Maier, O. 487
 Mailahn, N. 33, 233, 236, 255, 358-359
 Malle, B. F. 405, 407
 Marco, G. L. 404
 Martin, J.T. 393
 Mash, E.J. 218
 Mason, M.A. 481
 Masters, G. N. 174
 Matarazzo, J. D. 293, 399
 Maukisch, H. 478, 487, 500
 McBride, J. R. 393
 McClelland, M. C. 318
 McKinley, J.C. 36, 304, 398

- Meehl, P. E. 351, 354
 Mees, U. 97
 Meinefeld, W. 306
 Meyerhoff, H. 254
 Michel, L. 13, 19, 31, 41, 66, 71, 76, 94-96, 101, 104-105, 108, 123, 151, 154, 169, 174, 177, 179-181, 288, 319, 375
 Mische], W. 14, 16, 358
 Mittenecker, E. 300, 303-305
 Moser, K. 95, 149
 Müller, A. 464
 Müller, G.F. 376-377
 Mummendey, H. D. 299-305
 Münster, B. 308
 Murray, H.A. 14, 27, 107, 189, 211, 320, 325-326, 328, 455-456, 464
 Murstein, B.I. 319

 Nachreiner, F. 376-377
 Nanda, H. 85
 Neubauer, A. C. 405, 407
 Noelle, E. 219, 222, 236
 Nolen, P.A. 405
 Nordmann, E. 205
 Novick, M.R. 31, 71, 84, 91

 Olbrich, E. 358
 Odell, K. 405
 Olbrich, E. 33, 233, 236, 255-256, 359
 Ostendorf, F. 108, 111, 307

 Parry, H. J. 247
 Pawlik, K. 14, 96, 360
 Pelzmann, S. 426, 474
 Perez, M. 6
 Petermann, F. 4, 6-7, 9, 13, 16, 19, 198, 315, 338, 361
 Petermann, U. 7, 16, 315
 Pontalis, J.B. 187, 318
 Pryer, R.S. 319
 Pulver, U. 9, 441
 Quitmann, H. 15
 Raatz, U. 26, 38, 41, 54, 60, 62, 66, 78, 84, 87, 113, 119, 326, 346, 348, 408
 Rajaratnam, N. 85
 Rasch, G. 154, 174
 Rauch, M. 397
 Rauchfleisch, U. 218, 326
 Rausche, A. 301
 Reeb, W. 5, 212

 Rettig, K. 385, 392, 394-395
 Richter, H.E. 6-7, 53, 307, 427
 Ritz, B. 308
 Rock, D. L. 405
 Roest, F. 390
 Roidt, G.H. 38
 Ross, D. 190
 Ross, S.A. 190
 Ruch, W. 108, 111
 Ruffner, M. 415
 Rugg, D. 219
 Russell, J.T. 347
 Rutter, M. 250

 Saari, B. B. 108, 111
 Sarges, W. 219
 Sauermann, P. 266, 478
 Sawyer, J. 354
 Schaller, S. 14
 Scheiblechner, H. 174
 Scherer, K. R. 415
 Scheuch, E. K. 219, 222
 Scheurer, H. 360-361, 363
 Schmale, H. 282, 422
 Schmalt, H.-D. 422
 Schmid, F. W. 9, 441, 493
 Schmidt, G. 214
 Schmidt, K. J. 247
 Schmidt, L. R. 9, 247, 415, 441, 471, 473-474
 Schmidtchen, St. 9, 351, 373
 Schmidtke, A. 14
 Schmidtke, H. 422
 Schmitt, N. 108, 111
 Schmitt-Planert, A. 495-496, 498
 Schmitz-Scherzer, R. 255
 Schneewind, K. A. 188, 299, 309-310
 Schober, S. 317
 Scholz, O. B. 205, 218, 244, 314
 Schoppe, K. J. 68
 Schram], W. 199, 213, 247
 Schröder, G. 299, 309-310
 Schröder, R.D. 4
 Schuler, H. 358, 478, 495, 500
 Schulte, D. 245, 360
 Schwarzer, R. 315
 Schwenkmezger, P. 405-406, 409
 Seek, U. 247
 Sehringer, W. 330

- Seidenstücker, E. 247
 Seidenstücker, G. 247
 Seiffert, H. 414
 Seitz, W. 301, 351
 Selg, H. 36, 69, 183, 189-190, 454
 Sherman, R.E. 380
 Simon, T. 7, 422
 Sines, L.K. 247, 254
 Six, B. 306
 Sixtl, F. 27
 Skinner, B. F. 7
 Skre, I. 250
 Sloane, R.B. 370
 Solomon, J. 325
 Sommer, G. 52
 Sonnenberg, H.-G. 390, 401, 407
 Sonntag, K. 478, 481, 493-495, 498, 500
 Soskin, W. F. 254
 Spearman, C. 7, 267
 Spitznagel, A. 20, 188, 322-323, 417, 445
 Sprung, L. 10, 13, 19, 151, 180, 219, 338
 SPSS (Statistical Package for the Social Science) 385
 Staabs, G. von 204, 320, 331
 Staples, F.R. 370
 Starr, B. 325
 Stauffer, E. 315
 Steck, P. 326
 Stehle, W. 495
 Steinberg, M. 250, 256
 Stern, W. 6, 14
 Stevens, S. S. 28
 Stoll, F. 486-487
 Stone, M.H. 151-152, 154, 172-174
 Stults, D.M. 108, 111
 Stumpf, H. 34, 304, 383, 385, 387, 389, 392-396, 399, 404-407, 409
 Süllwold, F. 218
 Tanzer, N. 405
 Tatsuoaka, M.M. 309
 Taylor, H.C. 347
 Tent, L. 326
 Terdal, L. T. 218
 Testkuratorium der Föderation Deutscher Psychologenvereinigungen 407
 Tewes, U. 37, 119, 122, 264, 291, 299
 Thiel, R. 315
 Thomae, H. 6, 14, 186, 189, 211-212, 233, 255, 265, 441, 459, 461
 Thurstone, L. L. 7, 33, 36, 268, 271, 275
 Timaeus, E. 304-305
 Tinger, G. 399-400
 Tismer, K. G. 188, 199, 201
 Todt, E. 308, 422, 458
 Tomm, K. 219, 222, 245
 Tränkle, U. 31, 299-301, 310
 Trebeck, R. 478
 Trost, G. 7, 33, 359
 Trottmann-Geschwend, A. 315
 Truax, C.B. 370
 Turoff, M. 378
 Ueckert, H. 400
 Ulich, D. 213
 Undeutsch, U. 211, 236
 Urban, E. 405, 407
 Uslar, W. v. 427
 Vennen, D. 369, 503
 Völker, U. 15
 Vrana, S. 255
 Wagner, H. 44, 53, 198
 Wahl, D. 253, 308
 Waketield, H. 212
 Walbott, H. G. 415
 Walsh, V.R. 254
 Walsh, W.B. 254
 Walter, H. J. 437, 446, 469
 Wartegg, E. 330
 Warzecha, G. 385
 Weber, W. 397
 Wechsler, D. 67, 119, 122, 295, 463
 Wehner, E. G. 19
 Weidmann, M. 255
 Weinert, A. B. 480
 Weise, G. 5, 99, 124
 Weiss, D.J. 99, 391, 393
 Welsh, G. Sch. 36
 Wenninger, U. 386, 397
 Westhofen, R. 247-248, 250, 252, 255
 Westhoff, K. 9, 219-221, 227, 236, 247, 338, 441, 471
 Westmeyer, H. 4, 10, 341
 Weyerer, S. 250, 253, 255
 Whyte, W.F. 196
 Wieck, Th. 34

-
- | | |
|--|--|
| Wieczerkowski, W. 315 | Wurst, E. 171 |
| Wiggins, J. S. 297, 299-300, 351 | Wuttke, J. 24 |
| Wild, B. 393, 395 | Wyss, D. 15 |
| Wildgrube, W. 386-387, 395, 397, 405-406, 408-409 | Yager, G.G. 406 |
| Wilk, L. 214 | Yarrow, L. 218 |
| Wilks, S. S. 78 | |
| Williams, J. B. W. 250 | Zeidler, M. 483, 487 |
| Wilson, F.R. 406 | Zeisel, H. 236 |
| Windheuser, H. J. 370 | Zetterberg, H. 215 |
| Winslow, G.S. 250 | Zielinski, W. 4-6, 9, 19, 299, 315, 338, 344, 351, 441 |
| Wittchen, H.U. 250 | Zielke, M. 308 |
| Wittkowski, J. 227, 236, 321 | Ziler, H. 423 |
| Woodworth, R.S. 7 | Zinn, A. 198, 218, 238 |
| Wottawa, H. 6, 8-10, 19, 93, 111, 174, 176 178, 199, 300, 309, 338, 346, 369, 374, 377-380 | Zulliger, H. 323 |
| Wright, B. D. 151-152, 154, 172-174 | Zumkley, H. 320-321, 325, 327-328 |
| | Zuschlag, B. 338, 471 |

Sachregister

- Abweichungsnormen 114
 - ⇒ Variabilitätssnormen
- Act and React Testsystem (ART-90) 397
- Adaptive Tests 391
- Aktueller (u. biographischer) Ansatz 358
- Akzeptanz (einer Untersuchung) 423
- Alpha-Fehler (α) 344
- Analyse Qualitativer Daten (AQUAD) 397
- Anamnese 211-257
 - ⇒ Gespräch, Befragung, Exploration, Interview
- Angstfragebogen für Schüler (AFS) 315
- Anstrengungsvermeidungstest (AVT) 315
- Antwort, diagnostische 424-426
- Antworttendenzen (Fragebogen) 303-306
 - Ja-Sage-Tendenz 304
 - Kontrolle 210-213
 - Lügendenz 211
 - Simulations- / Dissimulationstendenz 304
 - Soziale Erwünschtheit 304
- APA-Normen 338
 - (APA: American Psychological Association)
- Äquivalenznormen (Normierung) 113
- Assessment-Center 491
- Aufgabe (Testaufgabe) ⇒ Item 35, 38-40
- Augenscheinvalidität 244
- Aufmerksamkeits-Belastungs-Test (Test d2) 269
- Axiome der klassischen Testtheorie (KTT) 70
- Bales: Interaktionsschema 198
- Basisrate 346
- BDP-Normen 338
 - (BDP: Berufsverband Deutscher Psychologen)
- Baum-Test 330
- Bedeutsamkeit (Meßtheorie) 28, 364
- Befragung 312-257
 - ⇒ Gespräch, Anamnese, Exploration, Interview
- Begutachtung 439-475
 - ⇒ Gutachten
- Behn-Rorschach-Test (BERO) 323
- Beobachtung 183-209
 - ⇒ Verhaltensbeobachtung
- Beobachtungsbogen für Kinder im Vorschulalter (BBK) 198
- Berufs-Interessen-Test II (BIT II) 308
- Berufsordnung für Psychologen 8-9, 337-339
- Berufswahltest (BWT) 397
- Beta-Fehler (β) 344
- Bewerber-Selektion 477-489
- Binet (Binetarium) 7, 113
- Binomialmodell 144
- Biographischer (u. aktueller) Ansatz 358
- Biographisches Inventar zur Diagnose von Verhaltensstörungen (BIV) 308
- Bühler-Hetzer-Kleinkindertest (BHKT) 270
- Children's Apperception Test (CAT) 325
- Complete Statistical System (CSS) 385
- Computer Adaptives Testen (CAT) 397
- Computordiagnostik 383-410
 - Computersysteme 388
 - Computertests 389
 - Adaptive Tests 391
 - Einsatzfelder 401
 - Äquivalenz zwischen Papier-Bleistift-Test und ihren Computer-Versionen 404
- Computergesteuertes diagnostisches System auf normativer Grundlage (DIASYS) 398
- Cut-Off-Point 141
 - ⇒ Kritischer Punktwert
- Datenschutz 337-339
- Decisionsstudie (D-Studie) 85

- ⇒ Generalisierbarkeitstheorie
- Delphi-Methode 378
- Deskription (Fragebogen, Persönlichkeits-test) 297-299
- Deskription (u. Performanz): Untersuchungsebene 357
- Dezentrales Testvorgabe- und Auswertungssystem (DELTA) 386
- Diagnostic and Statistical Manual of Mental Disorders (DSM-III-X) 401
- Diagnostik
 - Aufbau 19-22
 - Aufgabenfelder 6-7, 13-16, 273, 503-506
 - Definition, Sachbedeutung 3-4, 19-21
 - Entstehungsgeschichte 6-16
 - Ethischer, juristischer Kontext 8-10, 337, 415, 441
 - Materiale Diagnostik 20, 503
 - Modellvorstellungen 13-16
- Diagnostische Situation 20, 415
- Diagnostische Untersuchung 20, 413-429
- Differentieller Interessen-Test (DIT) 308
- Differentieller Wissenstest (DWT) 99
- Diskriminante (u. konvergente) Trennschärfte 52
- Diskriminante (u. konvergente) Validität 109
- Dissimulations- / Simulationstendenz 304
- L' Echelle Metrique de l'Intelligente (Binet & Simon) 7
- Eichstichprobe ⇒ Normstichprobe 111, 120
- Eigenschaft (trait) u. Diagnostik 14, 273
- Eignung 265
- Elektronische Datenverarbeitung u. Diagnostik
 - ⇒ Computerdiagnostik 383-410
- Entscheidungstheorie 373-381
- Erfolgskontrolle 369-372
- Ethical Principles of Psychologists and Code of Conduct (APA) 338
- Ethischer, juristischer Kontext diagnostischen Vorgehens 8-10, 337, 408, 415, 441
- Exploration 211-257
 - ⇒ Gespräch, Anamnese, Befragung, Interview
- Externale Konstruktionsstrategie 35
 - Test, Fragebogen
- Eysenck Personality Inventory (EPI) 112
- Fähigkeit 265
- Fehler der Informationsverarbeitung 199-201
 - ⇒ Verzerrungstendenzen
- Fertigkeit 265
- Formdeutungsverfahren 317-333
- Freiburger Persönlichkeitsinventar (FPI, FPI-R) 36, 305, 306
- Fragebogen 297-316
 - ⇒ Persönlichkeitstest
- Fragebogen zur Erfassung von Aggressivitätsfaktoren (FAF) 69
- Fragebogen zur Kontrollüberzeugung (IPC) 107
- Fremdbestimmung
 - ⇒ Selbstbestimmung 9-10
- Fuchs-Rarschach-Test (FURO) 323
- Geist-Bilder-Interessen-Inventar (GBII) 315
- Generalisierbarkeits-Studie (G-Studie) 85
 - ⇒ Decisionsstudie
- Generalisierbarkeitstheorie 85
- Generative Regeln (Kriteriumsorientierte Tests) 133
- Gespräch ⇒ Anamnese, Befragung, Exploration, Interview
 - Abgrenzung (Definition) 212
 - Arten / Klassifikation 214
 - Auswertung 236
 - Wiedergabe des Originalgesprächs 237
 - Zusammenfassung des Gesprächs 237
 - Schematische Z. 237
 - Thematische Z. 237, 239
 - Beitrag zu Diagnostik und Intervention 243
 - Durchführung 233
 - Fehler 247
 - Fragetechniken 219
 - Gütekriterien 247
 - Vorbereitung 226
- Gestalterische Verfahren 317, 330
 - ⇒ Projektive Verfahren
- Gießen-Test (GT) 51, 53, 307, 311
- Grundintelligenztest - Skala 3 (CFT 3) 99
- Grundkenntnisse, diagnostische 23
- Gruppentest für die soziale Einstellung (SET) 203
- Gutachten, psychologisches 439-475
 - Befund 456
 - Befundliste 462, 464
 - Befundskizze 463, 465

- Gliederung 441
- Richtlinien (BDP) 338, 439, 467
- Stellungnahme 466
- Tätigkeitsfelder 439
- Untersuchungsbericht 447
- Vorgeschichte 444
- Gütekriterien (Test) 66-110
- Objektivität 67
- Reliabilität 70
- Validität 93
- Guthnan-Skala 57
- Hamburg-Wechsler-Intelligenztest für Erwachsene (HAWIE, HAWIE-R) 37, 67, 122, 299
- Hamburger Neurotizismus- und Extraversionskala (HANES, KJ) 315
- Hamburger Persönlichkeitsfragebogen für Kinder (HAPEF-K) 53
- Hamburger Verhaltensbeurteilungsliste (HA-VEL) 198
- Heutismen (aus projektiven Verfahren) 324, 329, 332
- Hilfslosigkeits- und Selbstwirksamkeitsskala 315
- Hochrechnung 81-84
 - ⇒ Spearman-Brown-Formula
- Homogenität 54-59
- Faktorenanalyse 56
- Guttman-Skala 57
- Interkorrelation 55
- Rasch-Modell 59, 174
- Hypothesen 414
- Index der kategorialen Häufigkeit 41
 - ⇒ Schwierigkeit
- Integrative Diagnostik 411, 413-501
 - ⇒ Synthese in der Diagnostik
- Intelligenz(funktionen) 267
- Intelligenzmodelle 267
- Intelligenz-Struktur-Test (IST 70) 39, 69, 90, 93, 99, 393
- Interaktionistische Persönlichkeitspsychologie u. Diagnostik 16
- Internale Konstruktionsstrategie 36
 - ⇒ Test
- Interpretation 414
- Intervallskala 28, 123
- Intervention
 - Definition 5
 - Modelle 13-16
 - Beiträge zur Intervention
 - Klassische Testtheorie 122
 - Kriteriumsorientierte Testtheorie 148
 - Rasch-Modell 179
 - Verhaltensbeobachtung 203
 - Gespräch (Exploration, Interview, Anamnese) 243
 - Leistungstests 270
 - Persönlichkeitstest 313
 - Projektive Verfahren 321
 - Computerdiagnostik 401
 - Diagnostischer Prozeß 426
 - Therapieplanung 428
- Interview 211-257
 - ⇒ Gespräch, Anamnese, Befragung, Interview
- Intuitiv / rationale Konstruktionsstrategie 211-257
 - ⇒ Test, Fragebogen
- Item (klassische Testtheorie) 35, 38-40, 40-65
- Analyse (Itemanalyse) 40-65
 - Homogenität 54
 - Schwierigkeit 41
 - Trennschärfe 47
- Arten 38-40
- Generierung 38
- Selektion / Testrevision 59-65
- Item-Charakteristik-Kurve (ICC) 153, 175
- Itemparameter (Rasch) 151
 - ⇒ Personenparameter (Rasch)
- Juristischer, ethischer Kontext diagnostischen Vorgehens 8-10, 337, 415, 441
- Kinder: Gespräche mit K. 218
- Klassifikation (u. Selektion) 341-349
- Klassifikationsfehler 344
 - ⇒ Alpha-, Betafehler
- Klassische Testtheorie (KTT) 31-127
- Klinisch-psychologische / psychotherapeutische Intervention: Leitsätze zur Dokumentation (BDP) 338
- Klinische Urteilsbildung 351-355
 - ⇒ Statistische Urteilsbildung
- Kontrastfehler 201
- Konstruktionsstrategien (Test, Fragebogen)
 - (intuitiv / rational, external, internal)

- Konvergente (u. diskriminante) Trennschärfe 52
- Konvergente (u. diskriminante) Validität 109
- Konzentrations-tests 269
- Konzentrations-Leistungs-Test (KLT) 269
- Konzentrations-Verlaufs-Test (KVT) 269
- Korrespondenzprobleme (Untersuchung) 421
- Kriteriumsorientierte Leistungsmessung 129
- Kritische Differenz (zwischen Testscores) 92
- Kritik der klassischen Testtheorie 123
- Kritischer Punktwert (Cut-Off-Point) 141
- „Leerer Stuhl“ (Gestalttherapeutisches Verfahren) 438
- Leipziger Testsystem 396
- Leistung(sfunktionen) 266
- Leistungsmotivationstest (LMT) 315
- Leistungs-Prüf-System (LPS) 275, 393, 450
- Leistungstest 263-296
 - Allgemeine 267
 - Spezielle 269
- Lern- und Gedächtnistest (LGT 3) 68, 269
- Lincoln-Oseretzky-Scale, Kurzform (LOS KF 18) 270
- Logistische Funktion (Rasch-Skala) 154
- Lokale stochastische Unabhängigkeit 176
- Lügenskala 304
 - ⇒ Offenheit / Verslossenheit
- MAILBOX-90 390
- Messen 27-28
- Mildefehler 200
- Minderungskorrektur (Test, Validität) 102-105
 - Einfache 102
 - Doppelte 103
- Minnesota Multiphasic Personality Inventory (MMPI) 36, 306
- Modellvorstellungen (Diagnostik, Intervention) 13-17
- Multi-Attributive-Utility-Theory (MAUT) 172
- Multimethodale / Multimodale Diagnostik 411, 413, 431, 439, 477, 491
- Multiple choice (Mehrfachauswahl) 38
- Multitrait-multimethod-Validierung 108
- Nebengütekriterien 66
 - (Normierung / Ökonomie / Nützlichkeit / Vergleichbarkeit)
- NEO-Fünf-Faktoren Inventar (NEO-FFI) 307
- Nominalskala 28-29
- Normative Diagnostik 9
- Normen (Test) 111, 120
 - Äquivalenznormen 113
 - Eichstichprobe 120
 - kulturell-ethische Normen 122
 - Normalverteilung 120
 - Normieren 111
 - Probleme 120
 - Prozentränge 115
 - Rohwerte 111
 - Stichprobenabhängigkeit 121
 - Transformierte Werte 112
 - Übliche Normskalen 119
 - Variabilitätsnormen 114
- Normstichprobe 120
- Nutzen diagnostischen Vorgehens 373-382
 - ⇒ Utilität
- Object Relation Technique (ORT) 325
- Objektivität (Test) 66-69
 - Auswerterobjektivität 68
 - Durchführungsobjektivität 67
 - Interpretationsobjektivität 68
 - Probleme 69
 - Spezifische (Rasch-Skala) 177
- Offenheit / Verslossenheit 304
 - ⇒ (Fragebogen: Antworttendenz)
- Ordinalskala 28, 123, 179
- Papier-Bleistift-Tests und ihre Computer-Versionen 389
- Performanz (u. Deskription) 297-299
 - ⇒ Fragebogen, Persönlichkeitstest
- Personality Research Form (PRF) 33, 304
- Personenparameter (Rasch) 151-152
 - ⇒ Itemparameter (Rasch)
- Persönlichkeits-Entfaltungs-Verfahren 317-333
 - ⇒ Projektive Verfahren
- Persönlichkeitsinventar 317-333
 - ⇒ Persönlichkeitstest, Fragebogen
- Persönlichkeitspsychologie u. Diagnostik 13-16
- Persönlichkeitstest 317-333
 - ⇒ Fragebogen, Persönlichkeitsinventar
- Picture Frustration Test (PFT) 119, 326
- Plazierung 343

- Positionseffekt 200
- Probabilistische Testtheorie (Rasch) 151-182
- Profilanalyse (Testscores) 275
- Profilvergleich (Testscores) 282
- Projektion / projektiv 317-319
 - Identifikation 327
 - Klassifikation 319-320
- Projektive Verfahren 317-333
 - Umschreibung (Definition) 317
 - Arten / Klassifikation
 - Formdeutungsverfahren 322
 - Verbal-thematische Verfahren 325
 - Zeichnerische u. gestalterische Verfahren 330
- Beitrag zu Diagnostik und Intervention 321
- Probleme 320
- Prozentränge (Normierung) 115
- Prozeßdiagnostik 360
 - ⇒ Statusdiagnostik
- Prozeßorientierung u. Diagnostik 14
- Prüfsystem für Schul- und Bildungsberatung (PSB) 99
- Psychodynamischer Ansatz u. Diagnostik 15
- Rasch-Modell (probabilistische Testtheorie) 151-182
- Rationalskala 28-29
 - ⇒ Verhältnisskala
- Reagibilität 24
- Rechnergestütztes Psychodiagnostisches System (RPS) 396
- Reduzierter Wechsler-Intelligenztest (WIP) 37
- Reliabilität (Test) 72-92
 - Axiome der klassischen Testtheorie 70
 - Definition 72
 - Generalisierbarkeitstheorie 85
 - Halbierungsreliabilität 79
 - Hochrechnung 81
 - Inkompatibilität mit Validität 104
 - Konsistenz 86
 - Paralleltestreliabilität 78
 - Retestreliabilität 76
 - Spearman-Brown-Formula 81
- Reliabilitätsindex 101
- Rorschach 322
 - ⇒ Formdeutverfahren
- Rückmeldung 484
 - ⇒ Bewerberselektion
- School Apperception Method (SAM) 325
- Steno-Test 331
- Schwierigkeit(sindex: Test, Item) 41-46
 - Mehrstufige Antworten 43
 - Zweistufige Antworten 41
- Schweigepflicht 8-10, 337, 415, 441
 - ⇒ Vertraulichkeit 8-10, 337, 415, 441
- Sechzehn-Persönlichkeitsfaktoren-Test (16 PF) 299
- Selbstbeschreibung 297-299
 - (Fragebogen, Persönlichkeitstest)
- Selbstbeschreibung u. Verhalten 301-306
- Selbstbestimmung 9-10
 - ⇒ Fremdbestimmung
- Selektion (u. Klassifikation) 351-355
- Selektionrate 346
- Self-fulfilling prophecy 201
 - ⇒ Erwartung als Fehler
- Senior Apperception Technique (SAT) 325
- Simulations- / Dissimulationstendenz 304
- Sixteen Personality Factor Questionnaire (16 PF) 118
- Skalenniveau 28
- Spearman-Brown-Formula (Halbierungsreliabilität) 80, 83
- Social Desirability-Scale Edwards (SDS-E) 305
- Soziale Erwünschtheit (Antworttendenz) 304
- Sozialer Kontext der Diagnostik 8-9, 337-339, 415-418
- Split-half reliability 79
 - ⇒ Halbierungsreliabilität
- Standardisiert / Standardisierung 66
- Standardmeßfehler (Test) 90
- Standardschätzfehler (Test) 91
- Statistical Package for the Social Sciences (SPSS) 385
- Statistische (u. klinische) Urteilsbildung 391-355
- Statusdiagnostik 360
 - ⇒ Prozeßdiagnostik
- Synthese in der Diagnostik 411, 413-501
 - ⇒ Integrative Diagnostik
- Szenario-Technik 379
- Taylor-Russel-Tafeln 347
- Test
 - Arten 26, 37
 - Definition 24
 - Gütekriterien 66-108

- Itemanalyse 40-65
- Klassifikation 26, 37, 267-270
- Konstruktion
 - Klassische Testtheorie 31-127
 - Kriteriumsorientierte Leistungsmessung 129-150
 - Raschmodell 151-182
- Konstruktionsstrategien 35
- Kritik der klassischen Testtheorie 123
- Kritische Differenz 92
- Normierung 111
- Testrevision 59
- Vertrauensbereich 89
- Testbewertung
 - Kriterien 288
 - Beispiel (HAWIE-R) 291
- Test-Merkmal: Abgrenzung (Definition) 32
- Testtheorien
 - Klassische 31-127
 - Kriteriumsorientierte 129-150
 - Probabilistische (Rasch) 151-182
- Thematischer Apperzeptionstest (TAT) 325
- Thematischer Gestaltungstest (TGT-S) 325
- Trait u. Diagnostik 14, 273
 - ⇒ Eigenschaft
- Trennschärfe (Test, Item) 47-54
- Übersetzungsprobleme 418
 - (Proband / Untersucher)
- Ü-Koeffizient 135, 138
- Untersuchungsverlauf 411-429
- Urteilsbildung
 - Computer-gesteuert 401
 - Integration 424, 456, 466
 - Klassifikation 342
 - Selektion 245
 - Statistische, klinische 351
- Utilität diagnostischen Vorgehens 373-382
 - ⇒ Nutzen
- Validität 94-111
 - Abgrenzung (Definition) 94
 - Arten 95
 - Grenzen 101
 - Handlungs- 253, 255
 - Inkompatibilität (partielle) mit Reliabilität 104
 - Inhaltsvalidität / Kontentvalidität 95
 - Kommunikative 254
 - Konstruktvalidität 105
 - Konvergente / diskriminante 110
 - Kriteriumsbezogene 98
 - Minderungskorrektur (einfache / doppelte) 102-103
 - Multitrait-Multimethod-Validierung 108
 - Reliabilitätsindex 101
 - Übereinstimmungsvalidität 98
 - Vorhersagevalidität 100
- Variabilitätsnormen 114
 - ⇒ Normen / Normierung
- Variablen (Definition) 23
- Verbal-thematische Verfahren 325
- Varianten 325
- Verhaltensbeobachtung ⇒ Beobachtung
- Abgrenzung (Definition) 183
- Anwendung in der Diagnostik 203
 - (begleitend / selbständig)
- Beitrag zu Diagnostik und Intervention 203
- Einheiten bilden 187
- Einteilung 193
- Fehler
 - Allgemein 199
 - Speziell 201
- Fixierung (mechanisch / symbolisch) 197
- Gütekriterien 206
- Vorteile / Nachteile 206
- Veränderungsmessung 361
- Aporien 361
 - Bedeutsamkeitsproblem 364
 - Regression zur Mitte 365
 - Reliabilitäts-Validitäts-Dilemma
- Verhaltenstherapeutische Diagnostik: Kompendium 218
- Verhältnisskala 28-29
 - ⇒ Rationalskala
- Vertrauensbereich (eines Testscores) 89
- Vertraulichkeit 8-10, 337, 415, 441
- Verzerrungstendenzen 199-202
 - Erwartung (self-fulfilling prophecy) 201
 - Hofeffekt (halo effekt) 200
 - Kontrastfehler 202
 - Milde 200
 - Position 200
 - Strenge 200
 - Überstrahlungseffekt 200
 - Zentrale Tendenz 200
- Vier-Bilder-Test 492

Wartegg-Zeichen-Test (WZT) 330
Wiener-Testsystem 396
Wilde-Intelligenz-Test (WIT) 268

Zeichnerische Verfahren 317, 330
 \Rightarrow Projektive Verfahren
Zufallskorrektur (Schwierigkeitsindex) 42